Association analysis of genetic variants related to biotransformation, estrogen metabolism, and oxidative stress

A Calitz

orcid.org 0000-0001-6456-8090

Dissertation accepted in partial fulfilment of the requirements for the degree *Master of Science in Biochemistry* at the North-West University

Supervisor:	Dr G Venter
Co-supervisor :	Prof E Erasmus
Co-supervisor:	Dr M Schoonen
Assistant Supervisor:	Prof FH van der Westhuizen

PREFACE (AND ACKNOWLEDGEMENTS)

When venturing out on this journey of mastering oestrogen metabolism and an aspect in genetic variation, becoming a statistician was quite a surprise. A M.Sc. degree is a tough process that confronts a student to dedicate more time and thought to a project than ever before which is crucial for the advancement in the field. Yet I found myself enjoying exploring the working of the human body where simple systems working together to form a complex outcome. Merging multiple scientific fields together to better observe these complex outcomes is in essence what I believe biochemistry was from the beginning. The chemistry of the human body that is measured in physics determined parameters by mechanical engineered and IT specialist coded machines from biological collected samples. For my study by associating the metabolites measured from urine and serum in regard to biotransformation pathway and related-oxidative stress with genotyped SNPs from serum gDNA leads to new discoveries within healthy South African women that can be used as a reference when looking at disease models. Due to the use of genetics, oestrogen metabolomics and statistics many "common" principles are explained in regard to scientist who aren't adept in all three fields much like me at the start of this masters. The benefit is that most chapters, sections, tables and graphs can be viewed in isolation and still lead to an understanding of the topic in hand.

Acknowledgements:

- I am grateful to the CANSA organisation for funding the genotyping of this study's samples as well as a previous study that collected these samples.
- The Centre of Proteomics and Genetic Research (CPGR) in Cape Town who professionally genotyped this study's samples.
- I thank the North-West University for providing a post-graduate bursary that financed the education needed to complete this thesis.
- I show my appreciation to my supervisor and co-supervisors, Dr Gerda Venter, Dr. M Schoonen and Prof E. Erasmus whose guidance in each separate field was essential in the success of this study. The display of abundance patience will always be appreciated.
- I show my gratitude to my father who aided me in proof-reading the literature chapter of this thesis and emotionally assisting me in enduring the thesis writing period.

ABSTRACT

In a previous study by the NWU eBOSS research group, urinary oestrogen biotransformation metabolite levels and blood oxidative stress markers were measured in African and Caucasian premenopausal South African women. Variation in these metabolites was observed between individuals of the same as well as different ethnic groups. The aim of this hypothesis-generating study was to determine whether these variations correlated with genetic variants in certain biotransformation genes – namely *CYP1B1*, *CYP3A4*, *CYP1A1/2* and *COMT*. Genomic DNA was isolated from whole blood using a gDNA isolation with magnetic beads method. Thereafter samples were genotyped at the Centre for Proteomic and Genetic Research (CPGR) in Cape Town using the Global Screening Array (GSA) version 2.0 beadchip and the Illumina iScan. Due to small sample size, only SNPs for which there were at least 10 minor allele carriers (either heterozygous or homozygous for the minor allele), were selected in Haploview. PLINK logistics command highlighted specific SNP and metabolite interactions. Within the program SPSS, general linear models of univariate nature were implemented to determine the significance of these PLINK associations. 29 SNP-metabolite associations were determined to be significant where 15, five, five and four associations were found in Caucasian controls, African controls, Combined controls and Caucasian Combined oral contraceptive (COC) users respectively. Comparing Caucasian controls to African control, differences in metabolite levels and SNP frequencies resulted in no overlapping SNP-metabolite associations except for the Combined controls association being applicable to both population groups. The use of COC results in changes in metabolites levels and highlight new SNP-metabolite associations. Most of these associations have not been published before and may explain the variation in oestrogen metabolite levels and oxidative stress observed between individuals of similar or different ethnic groups.

KEYWORDS

Oestrogen, biotransformation, oestrogen metabolism, oxidative stress, breast cancer, *CYP*, *COMT*, South African, African, Caucasian.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

Abbreviations that are used more than once in this thesis are listed in this table.

Abbreviation	**Definition**
Ade	Adenine
BER	Base excision repair
BMI	Body mass index
Cluster Sep	Cluster Separation
COC	Combined oral contraceptives
COMT	Catechol-o-methyltransferase
CPGR	Centre of Proteomics and Genetic Research
CYP	Cytochrome P450 (e.g., *CYP1B1* – Cytochrome P450, family 1, subfamily B, Member 1)
ddNTP	Dideoxy nucleotides
DNA	Deoxyribonulceic acid
DRSP	Drospirenone
E_1	Oestrone
E_2	Oestradiol
E_3	Oestriol
eBOSS	Oestrogen Biotransformaiton and Oxidative Stress Status
EE	Ethinyloestradiol
ER	Oestrogen receptor
ERE	Oestrogen response element
ETC	Electron transport chain
FDR-BH	False discovery rate – Benjamini Hochberg

GC score	Gencall score
GLM	General linear model
GSA	Global screening array
GSH	Reduced glutathione
GSSG	Oxidized glutathione
GT score	Gentrain score
Gua	Guanidine
HSD	Hydroxysteroid dehydrogenase
HWE	Hardy-Weinberg equilibrium
LD	Linkage Disequilibrium
M	Major allele
m	Minor allele
MAF	Minor allele frequency
MB-*COMT*	Membrane bound -*COMT*
MM	Homozygous major allele genotype
Mm	Heterozygous genotype
mm	Homozygous minor allele genotype
NQO	Quinone reductase
NWU	North-West university
PLINK	Unknown
PR	Progesterone receptor
p-val	p-value
Q	Quinone
QO	Quality Control
ROS	Reactive oxygen species
RT	Room temperature

S-*COMT*	Soluble-*COMT*
SHBG	Sex hormone binding globulin
SNP	Single nucleotide polymorphism
SOD	Superoxide dismutase
SPSS	Statistical Package for the Social Sciences
SULT	Sulfotransferase
TF	Transcription factor

CHAPTER 1 – PROLOGUE

1.1 Introduction

Single nucleotide polymorphisms (SNPs) within DNA contributes to phenotypical diversity and further study on a molecular level thereof is necessary to reveal detail about pathway functions, gene-gene or gene-cofactors interactions and the aetiology of diseases such as cancer (Frazer *et al.*, 2009). These documented SNPs with a known function or effect are rarely directly associated with a specific metabolite measured within urine or blood (Thomas *et al*., 2011). The SNP-metabolite associations are measured per population or lifestyle choice since the factors such as SNP frequency, diet and lifestyle variously cause the phenotype changes seen between populations that are subsequently proven difficult to predict (Faber *et al*., 2005; Li *et al.,* 2014; Reding *et al*., 2012).

The North-West University (NWU), Oestrogen, Biotransformation and Oxidative Stress Status (eBOSS) research study is interested in the human hepatic capabilities to biotransform various substrates and to determine types and scale of pathway imbalances that could relate to specific diseases. As one of the five eBOSS sub-studies, this study's main focus is to generate SNP-metabolite (polymorphism with biotransformation) associations of premenopausal women per study group. These metabolites were acquired from the eBOSS research groups and were previously measured in serum and urine samples. The metabolites underwent a log transformation in order to achieve normal distribution necessary for association analysis. The genetic material was isolated by magnetic gDNA isolation attained from serum samples from the above-mentioned study and the purity, concentration and integrity quality was checked by spectrophotometry, fluorometry and electrophoresis-imaging respectively. The genetic material was sent to the Centre of Proteomics and Genetic Research (CPGR) for a wide range of specific chosen SNPs genotyping along with imaging. SNPs of the chosen genes of interest were selected by attaining the minimum requirements for association. Preliminary association analysis was done and yielded a summary of the metabolite and SNP selection. The time consuming, detailed association analysis was done and is the result presented in a summarized chart. The SNP-metabolite associations were visually displayed by boxplots in order to highlight differences between population groups or between both Caucasian controls and combined oral contraceptive (COC) users.

The statistical designated SNP-metabolite associations are listed and discussed per study group, per gene, per association and are unique to each study group. Key association results include increased or decreased metabolite levels and a minimum of at least 10 of 23-46 alleles samples in any category. External influences (cofactors) such as BMI, age, menstrual stages, and ethnicity was also reported when appropriate.

Comparing the associations between ethnicities or between both Caucasian controls and COC users could clarify the significant effect ethnicity, lifestyle (COC use) and genetics could have on metabolite levels. The larger quantity of oestrogen from the COC lifestyle choice could demonstrate an additional effect on oestrogen biotransformation pathways.

The samples acquired from healthy women can naturally demonstrate imbalanced pathways that can in time lead to the formation of serious life-threatening diseases (Singer *et al*., 2001). The change in metabolic levels along likely imbalanced pathway that associates to a SNP can lead to a better understanding of eventual disease aetiology and therefore improved treatment development or reveal possible preventative measures.

1.2 Problem Statement

Biotransformation is summarized as the process by which the liver can convert various substances into active or inactive states through a series of enzymatic reactions (Liska *et al*., 2006). These substances include foreign compounds (xenobiotics, e.g., preservatives from food or antibiotics) as well as endogenous compounds (endobiotics, e.g., hormones such as oestrogen) and are detoxified or degraded and excreted via this process (Liska *et al*., 2006). The biochemical reactions involved in this process are mediated by biotransformation enzymes and are mostly classified into two groups namely, phase I and phase II (Li *et al*., 2011; Liska *et al*., 2006). Phase I biotransformation reactions are catalysed by Cytochrome P450 (CYP) enzymes that add or expose specific functional groups while phase II reactions entail adding hydrophilic compounds to improve excretion. It can be concluded that the biotransformation system is very complex and consists of multiple phases which enables most healthy individuals to detoxify a wide variety of substances (Calitz *et al*., 2018). According to Li *et al*. (2011) and Liska *et al*. (2006), each enzyme in the biotransformation pathways have genetic polymorphisms such as single nucleotide polymorphisms (SNPs) that alter biotransformation enzyme activity and function or expression, resulting in individual metabolic differences. According to Li *et al*. (2011) information about biotransformation and excretion gene variations are very limited and is yet to be properly studied. Li *et al*. (2011) especially

stated that the phase I gene, *CYP3A4*, demonstrated highly differentiated genetics across African and non-African populations and even differences among closely related ethnic groups. This can be supported by the few South African studies published by (Aklillu *et al.,* 2002; Drögemöller *et al*., 2013; Swart & Dandara 2014) that yielded novel variants of the *CYP1B1*, *CYP1A2* and *CYP3A4* genes respectively. Sansom (2021) also adds to the conception that the South African populations especially the diverse African study groups have identified the most novel SNPs.

Furthermore, inter-ethnic genetic variation relating to oestrogen biotransformation pathways have been linked with an increased risk for oestrogen-induced breast cancer (Kato *et al*, 2009). Breast cancer (BCa) is the most frequent occurring type of cancer among women worldwide (Williams & Phillips, 2000). African women tend to develop a more aggressive type of BCa (hormone-receptor negative) which may be correlated with certain genetic variants (Kato *et al*., 2009). These may include genes associated with hormone biotransformation, e.g., *CYP1A1*, *CYP1A2*, *CYP1B1*, *CYP3A4* and *COMT* (Hodges & Minich, 2015; Guengerich, 2003). However, information about how oestrogen biotransformation is affected in different breast cancer subtypes (i.e., with different hormone receptor status) is scarce (Quan *et al*., 2014).

In a recent ongoing study performed at the NWU BOSS lab, oestrogen biotransformation, general biotransformation capacity, as well as oxidative stress status was measured. These measurements came from healthy African and Caucasian women between the ages of 18 and 35. Preliminary results from the eBOSS study indicate that there may be differences in phase I and phase II biotransformation capacities as well as in the capacity to regulate oxidative stress between individuals, but also between African and Caucasian women. Although these processes may be significantly affected by lifestyle factors, such as COC use, (Liska *et al*., 2006), the studies mentioned above suggest that genotype most definitely also plays a role.

Identifying specific genetic variants that determine biotransformation efficiency of hormones and other xenobiotics could aid in early detection of cancer risk. This, together with knowledge about how oestrogen biotransformation is affected can further shape the development of (preventative) treatment strategies in the form of nutritional regulation of biotransformation processes.

1.3 Aims and Objectives

This study has two main primary aims, that could both be achieved by following the list of objectives. The third aim, that is on a secondary level, compares the results achieved form the first two aims.

- Aim 1: To identify SNPs in selected biotransformation genes for association analysis from participants study groups of the eBOSS study.

- Aim 2: To determine whether the identified SNPs associate with increased or decreased levels of metabolites from the oestrogen biotransformation pathway and markers of oxidative stress in urine and blood.

- Aim 3: To compare the SNP-profile and the SNP-metabolite associations between the participant groups.

The objectives for aim 1 were to:

1) Identify participant study groups to be included.
2) Identify genes involved directly with oestrogen biotransformation and indirectly with markers of oxidative stress.
3) Isolate high quality gDNA (genomic DNA) from previously collected whole blood samples.
4) Generate high quality SNP-array genotypes for selected participants.
5) Identify high quality SNPs from selected genes for selected participant groups.
6) Select SNPs for downstream association analyses using Haploview.

The objectives for aim 2 were to:

7) Perform normalization of metabolite data for association analyses per participant group.
8) Perform preliminary SNP-metabolite association between SNP selected genes and normalised metabolite data in the four participations groups by using PLINK.
9) Identify three allele models to use in SPSS to evaluate possible SNP influence / effect on metabolite levels for final association analysis for the participant groups.

The objectives for aim 3 were to:

10) Compare association analyses profiles between different study groups using Minor allele frequencies (MAF) and Boxplots.

1.4 Basic Hypothesis

Some of the intra-ethnic and between inter-ethnic phenotypic variation in biotransformation and redox regulation of participants of the eBOSS study could be linked to differences in diet and lifestyles. We hypothesize that the remaining differences in biotransformation activity (such as methylation) of eBOSS participants will be explained by genetic variation in the genes associated with these processes.

CHAPTER 2 – LITERATURE REVIEW

Across the globe individuals express differences in phenotype. Although large parts of the genome are conserved among individuals, single nucleotide polymorphisms (SNPs) within certain genes make it possible to distinguish different genotypes and contribute to the diversity in phenotype. Differences in liver biotransformation efficiency as influenced by SNPs is one example of how genetic variation influences phenotype. Large variation is observed when comparing individual levels of metabolites derived from metabolic reactions occurring in the liver, including oestrogen biotransformation metabolites and by-products such as reactive oxygen species (ROS). These variations are of clinical interest since it could influence an individual's risk to develop diseases such as cancer, and also explain the differences in disease susceptibility between related individuals and between population groups. Although differences in disease susceptibility could be determined by environmental, dietary and lifestyle factors (e.g., the use of oral contraceptives), in addition to genetic composition, this study focussed on variations of the latter.

2.1 Liver Biotransformation Pathway

Numerous endogenous and exogenous compounds undergo a two-phase biotransformation in the liver either to be excreted, or transformed and further distributed (Liska *et al*., 2006). Multiple factors, including endogenous and exogenous molecules, regulate the activity of the enzymes in the above-mentioned pathway. These biotransformation enzymes are also expressed in some other tissue cells, and this could contribute to the measurable metabolite levels in urine and blood.

2.1.1 The two Phases of Biotransformation

The liver is an essential organ that executes approximately 500 major metabolite functions (Calitz *et al*., 2018). These hepatic cell functions include exocrine and endocrine secretions, protein biosynthesis and storage, biosynthesis of cholesterol, bile salts, and phospholipids, metabolizing carbohydrates and lipids, and biotransformation of endogenous and exogenous compounds (Calitz *et al*., 2018).

Smith and Williams (1970) first showed that non-reactive, lipophilic compounds undergo two phases of reactions during its biotransformation into hydrophilic compounds for excretion, as shown in Figure 2.1. Multiple referenced articles regarding the basic functioning and detailed description of biotransformation pathways have been reviewed by Liska *et al*. (2006) and is reiterated by most oestrogen metabolism and gene-specific articles listed within this study. The first phase was originally known as functionalization during which primary oxygen assists in the addition or revealing of a reactive site on the parent non-reactive compound. The second phase was known as conjugation that describes the process through which a water-soluble functional group is added to the reactive site facilitating excretion thereof in urine. In time, these two phases were renamed Phase I and Phase II detoxification.

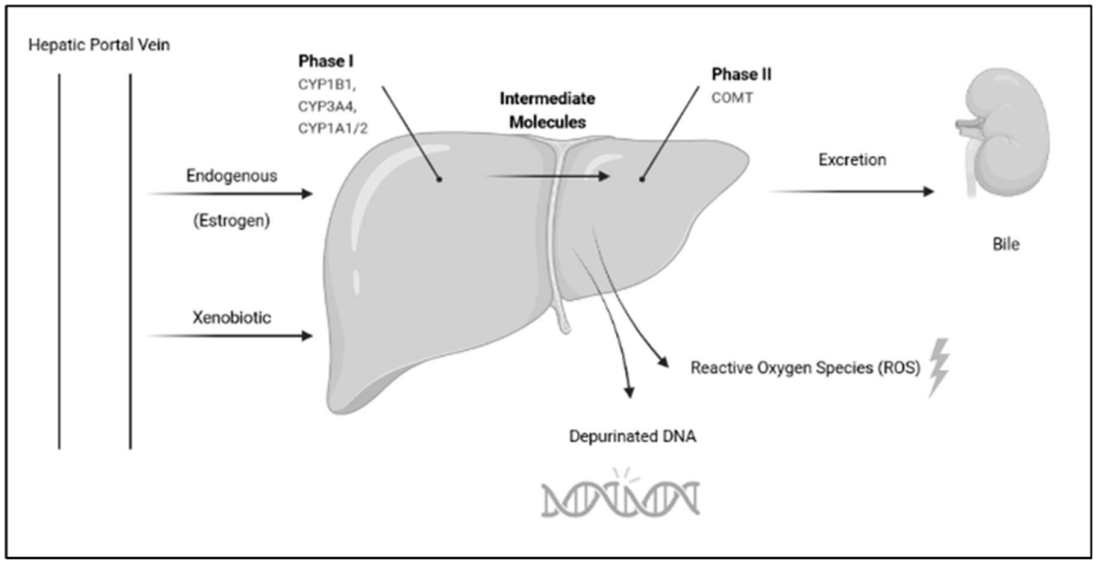


Figure 2.1 Liver Biotransformation Pathway

Legend

CYP – Cytochrome P450; *COMT* – Catechol-O-methyltransferase

Endogenous and xenobiotic metabolites are delivered to the liver by the hepatic portal vein. These metabolites are biotransformed in two phases by specific enzymes. After biotransformation, the metabolites are primarily excreted by the kidneys or liver bile but could also be reabsorbed by the circulatory system. For the purposes of this study, it must be noted that intermediate oestrogens could react with surrounding DNA after phase I to form DNA adducts that are mutation sensitive.

An abundance of Cytochrome P450 (*CYP*) enzymes catalyse phase I biotransformation reactions resulting in the production of transitional (intermediate) molecules. Hydroxylated metabolites, for example, are transitional molecules resulting from phase I reactions (Danielson, 2002). The non-reactive, parent molecules that undergo biotransformation were discovered to be either endogenously-produced hormones, signalling molecules (i.e., hormones with distant target cells), or exogenous molecules (also known as xenobiotics). Phase I reactions often bio-activate molecules into a more toxic or reactive state than that of the parent molecules. Consequently, Phase II reactions bio-inactivate these molecules. A critical balance exists between the two phases to prevent prolonged existence of the reactive intermediate molecules. Multiple enzymes, known as conjugases, catalyse phase II reactions by attaching molecules such as glucuronic acid, sulphate, specific amino acids including glycine and taurine, glutathione, or methyl groups to the intermediary molecule to increase water solubility. Some parent molecules are detoxified through biotransformation and excreted while others are metabolized into other signalling molecules, further demonstrating the diverse functions of these liver biotransformation enzymes. Optimal liver biotransformation (including xenobiotics) relies on a state of homeostasis between the two phases, with overloading or an imbalanced system resulting in an increased risk of cancer formation. The addition of antioxidants and specific nutritional support could, however, increase effective metabolite excretion.

For the purposes of this study, the following five specific genes considered prominent in phase I and phase II biotransformation processes were selected, namely (in order of chromosome number): *CYP1B1*, *CYP3A4*, *CYP1A1*, *CYP1A2*, and *COMT*. The *CYP* enzymes are part of a super family containing approximately 2 100 enzyme members (Alsubait *et al.,* 2020). *CYP*s regulate many human body biological processes that include biosynthesis and metabolism of oestrogen, conversion of cholesterol into bile acids, and bile acid biotransformation (Chen *et al*., 2014).

CYP1B1, forming part of the second chromosome pair of the typical 23 pairs, is known to have the longest mRNA sequence and number of amino acids of all the *CYP* enzymes, while also demonstrating a simplistic structure (Alsubait *et al.,* 2020). According to Alsubait *et al*. (2020), *CYP1B1* has three exons and two introns, and transcribes from the 5’ end of the second exon into 543 amino acids that could bind to a haem prosthetic group that catalyses oxidative reactions. *CYP1B1* plays an important role in the biotransformation of pre-carcinogens, such as polycyclic aromatic hydrocarbons (PAH) or certain oestrogen metabolites, by adding a hydroxyl group into the unreactive structures prior to phase II (conjugation) reactions (Danielson, 2002). It is important to note that *CYP1B1* is known to be expressed within the liver, as well as in other tissues (Alsubait *et al.,* 2020).

According to Katzung (2017), *CYP3A4*, a part of chromosome pair seven and one of the 42 *CYP* enzymes that metabolize xenobiotic and endogenous metabolites defined by Rendic (2002), represents 30% of *CYP* enzymes expressed in the liver (Katzung, 2017). *CYP3A4* is considered the most common xenobiotic detoxifier, metabolising approximately 50% of all the xenobiotics (Katzung, 2017). *CYP3A4* plays a major role in the activation of endogenous pre-carcinogens that includes bile acids and steroid hormones (e.g., testosterone and specific oestrogens) (Bai *et al.,* 2017). This versatile enzyme can be down regulated by the pregnane X receptor (*PXR*) in response to aryl hydroxylase receptor (AhR) that has been influenced by negative cross talk (Burton *et al.,* 2020).

CYP1A1 and *CYP1A2*, both part of chromosome pair 15, generally have the same biotransformation function. Expression of the *CYP1A* enzymes can be activated by the movement of an inducer AhR complex into the liver nucleus (Katzung, 2017, and Long *et al.,* 2017). *CYP1A2* is almost exclusively expressed within the liver (Faber *et al*., 2005), comprising approximately 15% of liver *CYP* enzymes (Katzung, 2017). *CYP1A2* is a rate-limiting step in the pre-carcinogen bioactivation of endogenous substrates (e.g., steroid hormones as mentioned for other genes in the previous sections) (Bai *et al.,* 2017). The role of *CYP1A2* in xenobiotic detoxification is specific to a few metabolites, and as such plays a lesser role in overall liver biotransformation (Faber *et al*., 2005). *CYP1A2* expression, activation, and inhibition are based on environmental, lifestyle, and genetic factors, with the consumption of multiple competitive substrates for *CYP1A2* considered the most prominent (Faber *et al*., 2005). The variability in *CYP1A2* and its specific xenobiotic interactions are still the subject of further study (Faber *et al*., 2005). It must be noted that *CYP1A2* yielded more relevant results within the framework of this study than *CYP1A1*.

According to Sak (2017), *COMT* is a dominant and essential Phase II enzyme. It forms part of chromosome pair 22, comprising six exons of which the first two are non-coding (Männistö & Kaakkola, 1999). *COMT* is a single gene containing two definite promoters within exon three that encode membrane-bound (MB-*COMT*) and soluble (S-*COMT*) *COMT* enzymes (Männistö & Kaakkola, 1999). Of these, S-*COMT* is abundantly expressed in most cells, whereas MB-*COMT* is expressed in smaller quantities (Männistö & Kaakkola, 1999). The highest concentration of *COMT* occurs in the brain, liver, and kidney (Sak, 2017). Intermediate molecules are generally rapidly inactivated by *COMT* and converted into non-toxic methylated metabolites, making *COMT* an efficiency measure of inactivation (Sak, 2017). The *COM*T enzyme is highly regulated by many factors and because of the very high capacity of even a few proteins it is not easily induced or inhibited (Männistö & Kaakkola, 1999). *COMT* catalyses inactivation of intermediate metabolites by facilitating the acceptance of a transferred methyl group from S-adenosylmethionine (SAM), resulting in the formation of a by-product named S-

adenosylhomocysteine (SAH) (Sak, 2017). In cases where increased SAH concentrations indicate a lack of methylated carriers, the SAH will act as a negative feedback inhibitor for the *COMT* gene (Sak, 2017). Additionally, the methylated metabolites produced by *COMT* could trigger a negative feedback loop that inhibits the hydroxylation activity of *CYP1A2* (Sak, 2017). However, it must be noted that exact Phase II mechanisms, including glucuronidation and sulphate conjugation, as well as the interaction between enzymes, have to date not been sufficiently studied (Männistö & Kaakkola, 1999). Most methylation, glucuronidation and sulphonation studies such as Tian *et al.* (2015) and Cisneros *et al* (2019) focus on a specific metabolite biotransformation and interactions in an attempt to clarify these mechanisms. *COMT* variation genotypes and its interaction with other enzymes could constitute a significant functional difference in metabolite levels (Sak, 2017). Männistö and Kaakkola (1999) note that the level of *COMT* activity differs between ethnic groups, with higher activity documented for African American versus that of Caucasian American population groups.

2.1.2 Biotransformation in Extrahepatic Cells

Biotransformation reactions do not necessarily occur only within the brain and liver but can also be encountered within other organs such as kidneys, stomach, and intestinal mucosal wall (Liska *et al*., 2006). *CYP3A4* catalytic activity and MB-*COMT* act as detoxification barriers between blood and other tissues within the intestine villi, and as such protect cells against harmful xenobiotic effects (Männistö & Kaakkola, 1999).

Biotransformation also occurs in extrahepatic cells where the expression of enzymes could differ from that within the liver (Badawi *et al*., 2001). *CYP1A2* and *CYP3A4* enzymes are typically well expressed in hepatic cells leading to increased pathways products (Badawi *et al*., 2001; Moon *et al*., 2006). *CYP1A1* and *CYP1B1* enzymes are more dominantly expressed in extrahepatic cells (Moon *et al*., 2006). In the latter, breast tissue is noted to express high levels of *CYP1B1* enzymes considered reactive intermediate metabolites (Samavat & Kurzer, 2015). Discovery of the above-mentioned differences in enzyme expression, and as such the dominant pathways active within different tissue types, facilitates understanding of the aetiology associated with certain tissues, e.g., breast cancer (BCa).

2.2 Oestrogen Metabolism

Oestrogen is a collective name for multiple steroid hormones that are synthesized within ovaries and other peripheral (i.e., lesser central) tissues (Samavat & Kurzer, 2015). The main

functions of these oestrogen hormones are to regulate growth and the human reproductive systems (Lee *et al*., 2012a). The complex interaction between the different structures and activities of these oestrogen metabolites are best described by means of the oestrogen biosynthesis process that produces three main oestrogen metabolites. These primary oestrogens are oestrone (E_1), oestradiol (E_2) and the 16-hydroxyoestradiol designated as oestriol (E_3) (Samavat & Kurzer, 2015). The primary function of E_1, being quite abundant but less potent oestrogen, is to regulate reproductive systems and to store oestrogen by enzymatically catalysed sulphate binding (Mauvais-Jarvis *et al*., 2013; Samavat & Kurzer, 2015). E_2 more actively regulates the development and maintenance of reproductive and cardiovascular systems (Zou & Ing, 1998). Additionally, E_2 is the most biological active and abundant oestrogen in premenopausal women (Samavat & Kurzer, 2015). As E_3 plays a key role during pregnancy, its presence is expected to be elevated in this period (Falah *et al.,* 2015). These oestrogen metabolites affect a wide range of target tissues and are eventually transported to the liver through the blood circulatory system for biotransformation (also known as oestrogen metabolism). Oestrogen metabolism occurs along numerous pathways exhibiting different prevalences and impacts on the surrounding environment or target cells. Multiple genes are involved in maintaining the balance within the pathways. An increase in oestrogen metabolites due to exogenous consumption could amplify certain pathway metabolites and its associated impacts.

2.2.1 Oestrogen Biosynthesis and Transportation

This section provides a summary of oestrogen biosynthesis as described primarily by Samavat and Kurzer (2015) and Saini *et al*. (2021). All steroid hormones originate from C-27 cholesterol, itself mainly derived from LDL-cholesterol. C-27 is biotransformed into several products such as C-19 steroid (androstenedione) that acts as a the biologically inactive androgen precursor for both E_1 and testosterone (E_2 precursor) catalysed by 17β -*HSD* (hydroxysteroid dehydrogenase) within the ovaries and peripheral tissues such as adipose tissue. Testosterone is converted to E_2 by the rate-limiting enzyme Cytochrome P450 aromatase (*CYP19A1*) present in peripheral tissue of target tissues. Although not the primary pathway, testosterone can be deactivated within the liver by the *CYP1B1* and *CYP3A4* enzymes (Beuten *et al.,* 2008; Qian *et al.,* 2017). Further metabolizing of C-19 steroids produces C-18 steroids consisting of a benzene ring, phenolic hydroxyl group at C-3, and either a ketone (E_1) or hydroxyl (17ß-E_2) functional group at C-17, making oestrogen one of the few aromatic molecules in humans.

According to Ozougwu (2017), non-bounded fractions of oestrogen hormones circulating in the blood stream carried by sex hormone-binding globulin (SHBG), a glycol protein produced by the liver, can possibly through the hepatic portal vein interact with liver cells (Figure 2.1). As such, SHBG is relevantly upregulated by testosterone and oestrogen hormones (Rettberg *et al.,* 2014). Prior to oestrogen metabolism in the liver, E_1 can be reversibly converted to E_2 through an enzyme encoded by the *HSD17ß1* gene, although this enzyme has other primary functions in other target tissues (Samavat & Kurzer, 2015). The primary oestrogen hormones, biosynthesized as described in this section, are mostly stored as E_1-sulphate or E_2-sulphate (also known as E_2-3-sulphate) (Hobkirk, 1993).

2.2.2 Oestrogen Metabolism Phases and Pathways

An overview of the oestrogen metabolism process based on work conducted by several authors (e.g., Cavalieri& Rogan, 2014; Janacova, 2015; Samavat & Kurzer, 2015; Sood *et al*., 2017 and Trabert *et al.,* 2016) is provided in diagrammatical format in Figure 2.2.

Since oestrogen hormones are metabolized within the liver, it follows the same biotransformation phases as described in Section 2.1.1 (noted on the left of Figure 2.2). Phase I metabolism of oestrogen hormones occurs along three major pathways resulting in the irreversible formation of 2/ 4/ 16α-hydroxylated oestrogens collectively known as catechol oestrogens (CE) which are deemed reactive intermediate metabolites (Cavalieri & Rogan, 2014). These reactions are catalysed by competitive, NADPH-dependent enzymes encoded by *CYP* genes associated with Phase I reactions.

The major pathway in this process is 2-hydroxylation of E_1 and E_2 that predominantly produces the less carcinogenic 2-hydroxyoestrogens (Samavat & Kurzer, 2015). More specifically, *CYP1A1/ CYP1A2/ CYP3A4* catalyses hydroxylation of E_1 and E_2 at preferably C-2 position within the endometrium and hepatic cells (Ashton *et al.,* 2009; Long *et al.,* 2007). It is important to note that the *CYP1A2* enzyme has a higher affinity for E_1 (Dumas & Diorio, 2011). These 2-hydroxylated metabolites comprise approximately 36% of oestrogens (Eliassen *et al.,* 2012). These are, however, considered to have much weaker activity within the target organs or receptors than both the parent molecules (Sak, 2017).

The *CYP1B1* enzyme only catalyses the least favoured hydroxylation phase I pathway at the C-4 position of both E_1 and E_2, occurring more frequently within extrahepatic cells (Dumas & Diorio, 2011). The resultant 4-hydroxylated oestrogens possess a similar potency in activity as its parent molecules and are therefore found in smaller quantities (i.e., only approximately 4% of oestrogens) (Eliassen *et al.,* 2012; Kato *et al.,* 2009).

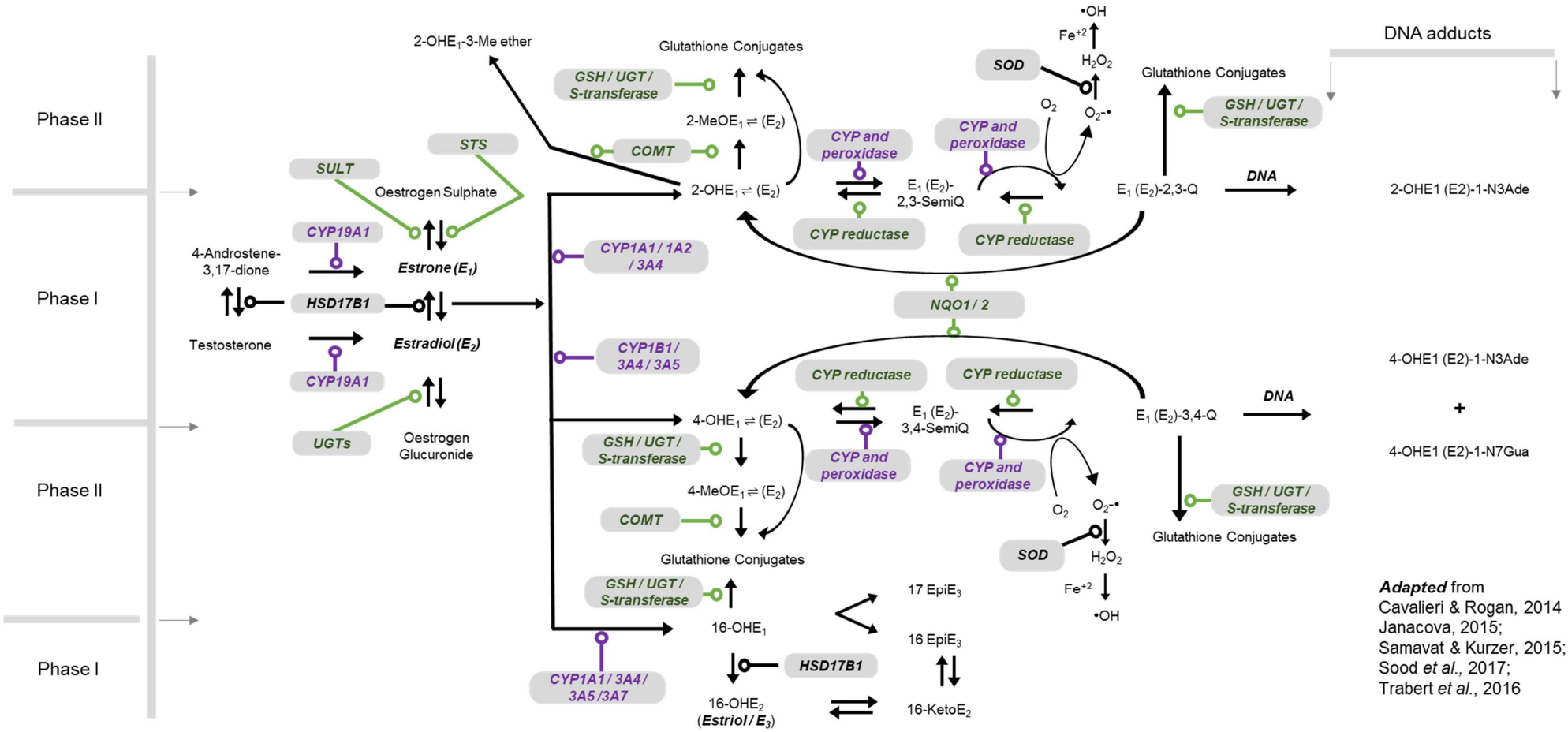


Figure 2.2 Oestrogen Metabolism

Legend
OH – Hydroxy; Me – Metho (Methyl); Q – Quinone; N – nitrogen; Ade – Adenine; Gua – Guanine;
CYP – Cytochrome P450; *HSD* – Hydroxysteroiddehydrogenase; *SULT* – Sulfotranferases; *STS* – Steroid sulfatase; *SOD* – Superoxide dismutase;
COMT – Catechol-o-methyltranferase; GSH – glutathione-S-transferase; *UGT* – UDP-glucuronosyltransferase;

All enzymes within the figure are highlighted. The enzymes shown in green are carcinogenic protective while the purple enzymes are DNA-adduct inducers

The third pathway during phase I oestrogen metabolism produces 16α-hydroxylated-E_1 and 16α-hydroxylated-E_2. The 16α-hydroxylated-E_1 can be converted to E_3 when necessary (Samavat & Kurzer, 2015), being the most abundant in pregnant women and the second highest oestrogen metabolite level in urine comprising 38% of all oestrogens (Eliassen, 2012).

The CE from phase I can be deactivated during phase II conjugation of glucuronide or sulphate, especially in the liver, or to a methyl group especially in extrahepatic cells (Cavalieri & Rogan, 2014). The glutathione (GSH) is conjugated to CE by glutathione-S-transferase (*GST*) or UDP glucuronosyltransferase (*UGT*), while sulphate conjugation is catalysed by S-transferase, and the methylation of the 2- / 4-CE is catalysed by catechol-O-methyltransferase (*COMT*) (Samavat & Kurzer, 2015; Hodges & Minich, 2015). All the inferred reactive intermediate CEs could be transformed into stable and non-toxic methoxy-oestrogens (Sak, 2017). The function of these methoxy-oestrogens differs significantly from that of the parent molecules (Männistö & Kaakkola, 1999).

Additionally, all CE metabolites, especially 2-hydroxylated E_2, are high affinity-substrates for *COMT* and could oversaturate the *COMT* enzyme when present in very high concentrations (Männistö & Kaakkola, 1999; Sak, 2017). In this light, methylation by *COMT*, rather than glucuronidation or sulphonation, plays the biggest role in efficient CE conjugation deactivation (Sak, 2017). After methylation, the conjugated metabolites are usually excreted by the kidneys to form part of urine (Dumas & Diorio, 2011). Sak (2017) notes that the ability to eliminate CE is dependent on hereditary material (i.e., genetics). It is important to note that E_2 can be reversibly oxidised into the E_1 form during either phase I hydroxylation or phase II methylation processes (Männistö & Kaakkola, 1999).

When not deactivated by conjugation, the CE could follow a separate pathway along which it is converted into semi-quinone and quinone oestrogens that could adversely interact with DNA by the formation of superoxide radicals resulting from a redox cycle between the CE and quinones (Sak, 2017). In addition, all quinones could irreversibly bind to DNA to form DNA-oestrogen-adducts, with 2-hydroxyoestrogen quinone forming stable 2,3-DNA-adducts (Sak, 2017; Männistö & Kaakkola, 1999). However, the 4-hydroxyoestrogen produces more reactive semi-quinone and quinone metabolites that are more prone to the formation of unstable, depurinating 3,4 DNA-adducts (Sak, 2017; Dumas & Diorio, 2011; Männistö & Kaakkola, 1999). Any inhibition of the normal metabolism of oestrogen could lead to an increase in 4-hydroxylation and as such an abundance of 3,4 DNA-adducts (Männistö & Kaakkola, 1999). A purine base is typically removed during the formation of these DNA-adducts, rendering these sensitive to single base mutations or genomic deletions which could trigger the start of cancer formation (Sak, 2017). Normally, the CE quinone pathway is disrupted by a set of activation and protective enzymes that sustain homeostasis, where the resulting low or depleted levels

of CE quinones thus minimize interaction between reactive metabolites and DNA (Cavalieri & Rogan, 2014).

2.2.3 Importance and Function of Oestrogen Metabolites

In addition to the previous generalized discussion regarding oestrogen biosynthesis, phase I and II metabolism, and quinone formation, the importance and function of the various metabolites (depicted in Figure 2.2) require further explanation.

The produced E_1, as well as E_3 resulting from either the reduction of 16-hydroxylated E_1, or 16-hydroxelated E_2, are known to extensively circulate to a variety of cells in the body (Janacova, 2015). Given that E_1 exhibits lower oestrogenic activity than E_2, E_1 readily converts into E_2 through the 17ß-*HSD* enzyme even after hydroxylation or methylation in response to regulatory triggers in the body (Janacova, 2015). An excess of E_1 and E_2 is sulphonated through conjugation by *SULT1A1, SULT2A1,* and *SULT1E1* enzymes to be temporarily stored before metabolism (Calderón, 2022).

E_2 is converted into 2-methoxy-oestrogens after hydroxylation by specifically *CYP1A1* or *CYP1A2* at the C-2 position (Sepkovic & Badlow, 2009). As previously mentioned, *CYP1B1* catalyses 4-hydroxylation of oestrogens, leading to the formation of 4-methoxy-oestrogens (Sepkovic & Badlow, 2009). *CYP1A1*, *CYP1A2*, and *CYP3A4* could also hydroxylate oestrogens at the C-4 position, although the activity of these enzymes decreases 12-,5-, and 9-fold, respectively, when compared to that of *CYP1B1* in this specific reaction (Badawi *et al*, 2001).

The 2-hydroxyoestrogens influence multiple pathways in an anti-oestrogenic manner since these hydroxylated metabolites exhibit lower binding affinity to ER, lower activity, and lower potency than that of the parent molecule (Männistö & Kaakkola, 1999; Samavat & Kurzer, 2015). More specifically, the 2-hydroxylated oestrogens inhibit cell proliferation and growth (Samavat & Kurzer, 2015). 2-hydroxy-E_1, E_3 and E_1 comprise the most abundant oestrogen metabolites circulating through the cardiovascular system, representing approximately 27, 18 and 15% thereof, respectively (Eliassen *et al.,* 2012).

Conversely, not all CEs demonstrate an overall anti-oestrogenic effect, since 4-hydroxy-E_2 exhibits a higher and longer-lasting binding affinity to ER with similar hormonal activity and potency than the parent molecule that is associated with various cancers (Männistö & Kaakkola, 1999; Cavalieri & Rogan, 2014). In this light, 4-hydroxyoestrogens is considered to be involved with the aetiology of numerous cancers (Westerlind *et al*., 2000).

16-hydroxyestradiol (E_3), being quite abundant, serves as a pathway aiding the excretion of E_1 and E_2 (Eliassen, 2012). The hydroxylation on the C-16 position of E_1 and E_2 is specifically

catalysed by *CYP3A4* or *CYP3A5* enzymes, with these oestrogens acting as the precursors for not only E_3 but also 16-Keto-E_2 and Epi-oestrogens, amongst others (Sepkovic & Badlow, 2009). The results of breast cancer (BCa) studies reveal that E_3 can be considered a protective metabolite that acts as an inhibitor to cancer-inducing metabolites. (Gorbach, 1984).

2- and 4-methoxy-oestrogens from phase II oestrogen metabolism are the least abundant of the oestrogen metabolites (comprising between 1 and 5%) and as such have been the focus of only a few studies (Eliassen *et al.*, 2012). Männistö and Kaakkola (1999) state that methoxy-oestrogens exhibit minimal to no affinity to oestrogen receptors (ERs) and subsequently does not demonstrate oestrogenic effect on oestrogen-sensitive tissues. Its' higher affinity for SHBG compared to that of E_2 allow 2-methoxy-oestrogens to readily distribute along the cardiovascular system (Dubey & Jackson, 2001). Of the 2-methoxy-oestrogens, methoxy-E_1 and the precursor metabolite, 2-hydroxy-E_1, are more abundant in plasma and urine (Männistö & Kaakkola, 1999). The less mentioned 4-methoxy-oestrogens exhibit slightly lower activity than its 2-methoxy-counterparts (Männistö & Kaakkola, 1999).

CE conjugation occurring either with, or without methylation, also forms part of the phase II oestrogen metabolism. This is generally achieved by multiple liver *SULT* and *UGT* enzyme families, with E_2 conjugating into E_2-3-glucuronide and E_2-17-sulphate (Raftogianis *et al.,* 2000; Calderón, 2022). Of these conjugated CEs, the E_1-sulphate metabolite is measured in larger quantities than E_1 present within the circulatory system (Raftogianis *et al.,* 2000). However, sulphate conjugation inactivates the CE, allowing its use as storage medium until transportation across a membrane causes the sulphate group to detach with subsequent reactivation of the CE (Raftogianis *et al.,* 2000). In particular, 16-hydroxylated metabolites undergo conjugation with sulphate and glucuronide as a second phase to oestrogen metabolism (Jiang *et al.,* 2009). This results in the formation of four main E_3 conjugates, namely E_3-3-sulphate, E_3-16-glucuronide, E_3-3-glucuronide and E_3-3-sulphate-16-glucuronide, with the first two products known to conjugate within the liver (Jiang *et al.,* 2009; Gorbach, 1984). The E_3-16-glucuronide is either excreted into the urine or converted into the double conjugate E_3-16-glucuronide-3-sulphate before being released into the bile (Gorbach, 1984). Although oestrogens containing glucuronide is predominant in bile, the double conjugate E_3 is the most abundant constituent thereof (Bolt, 1979). The intestinal bacteria deconjugates conjugated oestrogens, reverting these into CE forms that are mostly re-absorbed (Gorbach, 1984).

In literature, emphasis is predominantly placed on E_3 or E_3-glucuronide conjugates, while less is known about E_3-3-sulphate. Regarding the latter, Dawson *et al*. (2015) state that sulphonation of steroid hormones adjusts the bioactivity thereof and as such can detoxify xenobiotics (e.g., contraceptives). However, as the main functions of E_3 and E_3-sulphate

conjugates are related to pregnancy (Dawson *et al*., 2015), it is not in the interests of this study to discuss these metabolites in more detail.

An alternative pathway that occurs between the phases as shown in Figure 2.2 can form ROS (Li *et al*., 2017). The intermediate CE metabolites could be catalysed into semi-quinone and thereafter into quinone oestrogens (e.g., stable 2,3-E_2-quinone or unstable 3,4-E_2-quinone) through reduction and oxidation cycles, producing ROS as toxic by-products (Li *et al*., 2017). ROS can be detoxified as a result of scavenging by antioxidant enzymes, such as superoxide dismutase (SOD) (Mittler, 2017). However, a build-up of ROS, known as oxidative stress, could occur when ROS production exceeds scavenging ability, causing damage to surrounding biological molecules (especially protein, lipids and DNA) through oxidation that overwhelms damage repair systems (Li *et al*., 2017; Mittler, 2017). It is important to note that damage to DNA could initiate cancer formation (Ray *et al.,* 2012). That said, the highly reactive 4-hydroxylation pathway causing the formation of the most quinones is also the least favoured, and as such typically produces relatively low ROS levels (Dumas & Diorio, 2011; Li *et al*., 2017).

During interaction between DNA and oxidised CE, both semi-quinone and quinone oestrogens could covalently bind to a DNA purine base on the N-3 position of adenine or the N-7 position of guanine, defining DNA-oestrogen adducts (Janacova, 2015). The ability of 2-hydroxylated oestrogen quinones to damage DNA differs from that of 4-hydroxylated oestrogen quinones, as the former form reversable and stable DNA-adducts (Eliassen *et al*., 2012). In contrast, the 4-hydroxylated semi-quinone and quinone oestrogens cleave the glycosidic bond between the nitrogenous base and deoxyribose sugar, thereby removing the purine base from the DNA structure (Janacova, 2015). This creates an apurinic site sensitive to single strand breaks that is prone to mutations along with an unstable DNA-oestrogen adduct (e.g., D-4OHE_2-1-N7G) (Janacova, 2015). Additionally, an apurinic site that lacks possibly important genetic information can be passed on during DNA replication, while unstable DNA-oestrogen adducts could randomly chemically react with other molecules (Eliassen *et al*., 2012; Nakamura *et al*., 1998). A balanced oestrogen metabolism mostly maintains low DNA-oestrogen adduct levels that occasionally discard into the blood stream after formation of apurinic sites on the DNA (Pruthi *et al*., 2012). However, these apurinic sites can be repaired by apurinic/apyrimidinic (AP) endonucleases repair enzymes through a process called base excision repair (BER), although an imbalance between the number of apurinic sites and expressed repair enzymes can increase mutation opportunities (Nakamura *et al*., 1998; Sak *et al*., 2017).

2.2.4 Product and by-product Excretion and Removal

Oestrogen metabolites detoxified into water soluble metabolites during phase II conjugation form glucuronide, sulphate or methyl-conjugated oestrogens hydrophilic enough to be swiftly excreted by the kidneys to form part of urine (Dumas & Diorio, 2011; Liska *et al*., 2006; Samavat & Kurzer, 2015). In some instances, oestrogen metabolites biotransformed during 16-hydroxylation-specific phase II conjugation exhibit lipophilic properties and are mostly excreted into bile.

E_1 and E_2 utilizing the 16-hydroxylation pathway can be converted into double conjugate E_3-16-glucuronide-3-sulphate with increased half-life that released into the bile and as such could either act as signalling metabolites or is excreted (Bolt, 1979; Gorbach, 1984; Samavat & Kurzer, 2015). The conjugated oestrogen within the bile is regularly deconjugated by intestinal bacteria and subsequently re-absorbed, which leads to the observation that only half of the bile oestrogens are typically conjugated and only approximately 10% is eventually excreted (Gorbach, 1984).

Additionally, DNA-oestrogens adducts, another by-product of oestrogen metabolism resulting along these pathways, are discarded into the circulatory system and eventually excreted by the kidneys into urine (Pruthi *et al*., 2012).

2.2.5 Exogenous Oestrogen: Combined Oral Contraceptives

The liver is considered an essential organ responsible for the metabolism of large quantities of almost an infinite range of xenobiotics (Calitz *et al*., 2018). This function could be influenced by the hepatic blood flow, xenobiotic binding-affinity to plasma-protein, hepatocyte adsorption, the cell wall integrity, and the regulated state of the bile formation system (Calitz *et al*., 2018). The xenobiotics known as oral contraceptives are intermediate to very lipophilic compounds and as such are more difficult for enzymes to access (Liska *et al*., 2006). Therefore, the rate of biotransformation is reduced, resulting in longer lasting effects associated with these xenobiotics (Calitz *et al*., 2018). Combined oral contraceptives (COC) also regularly interact with other xenobiotics delaying biotransformation of both, resulting in increased burden on the liver while decreasing pharmacological effects of the other medication (Zhang *et al*., 2007). In this light, it is evident that COC are less susceptible to metabolism within the liver (Naz 2014).

Intake of COCs, is the most effective and common method for the prevention of pregnancy but only when adhering to the correct consumption schedule (Anzai *et al*.,2012). COCs generally comprise a combination of progestin (synthetic progesterone that causes less side-effects than androgens) and synthetic oestrogen (De Leo *et al*., 2016). The primary role of COC is to intercept ovulation through inhibition of gonadotropins [e.g., luteinizing hormone

(LH)], while progestin also decreases the endometrium receptivity (Naz, 2014). Naz (2014) states that the success of COC is between 95 and 99.8% as determined on a yearly basis. Although COC are helpful, women casually neglect to adhere to the consumption schedules or abruptly discontinue the use thereof leading to greater adverse effects and tolerability deviations than normally expected (Anzai *et al*.,2012). Therefore, over time COC have been adjusted to remain effective, reduce adverse effects and increase bodily acceptance of the product (Anzai et al.,2012).

The two COC of interest to this study, commonly known by the brand names YASMIN and YAZ, both consist of ethinyloestradiol (EE) and drospirenone (DRSP), the latter being a fourth-generation synthetic progestin (Anzai et al.,2012; Li & Anderson, 2010; Naz, 2014). EE reduces LH and follicle-stimulating hormone (FSH), with the former regulating steroid hormone production and the latter inhibiting egg growth inside ovaries (Naz, 2014). DRSP has antiandrogenic properties (i.e., it inhibits the binding of androgen to a specific receptor) and as such prevents androgen from regulating gonadotropins (Naz, 2014). YASMIN is associated with satisfactory tolerance against adverse effects, as well as proper cycle regulation (Anzai *et al*.,2012). Li and Anderson (2010) state that only YAZ could improve physical and emotional symptoms in women who regularly experienced menstrual irregularities. These two specific COC were specifically formulated to improve the function while limiting side effects exhibited by previous products (Krauss & Burkman,1992; Van der Meer*,* 2017). It is important to note that the oestrogen and progestin contained within these COC are reported to still affect the same multiple metabolic pathways and cause inflammatory reactions (Krauss & Burkman, 1992). As an example, Rodrigues (2022) found that COC offset oestrogen metabolism balance by increasing phase II enzyme activity and inhibiting phase I enzymes. This imbalance could limit produced glutathione-conjugated metabolites, thereby decelerating many essential processes that involve this metabolite (Jumuddin, 2018). In general, oral contraceptives have been linked to significantly decreased levels of testosterone and increased levels of SHBG which binds and limits E_2 and testosterone target cell interactions (Skibola *et al.,* 2005).

COC not only influence metabolites but also genetic activity and expression of oestrogen metabolism enzymes (Zhang *et al.* 2007). Since COC are similar to bodily produced oestrogens, they contain carbon chains and exhibit aromatic properties and can be readily hydroxylated by *CYP* enzymes being the first safeguard against adverse effects exhibited by COC (Danielson, 2002; Hodges & Minich, 2015). Occasionally, *CYP* hydroxylation could lead to destabilization of a functional group in EE that could covalently bind to the *CYP* enzyme thusly inhibiting further catalyses (Zang, 2007). However, during normally occurring EE metabolism 2-hydroxylated EE most likely undergoes glucuronide or sulphate conjugation by *UGT* and *SULT* enzyme catalyses (Zang, 2007).

2.3 ROS derived from Oestrogen Pathways

The results of ongoing metabolic research by other researchers at the North-West University indicate that the levels of ROS significantly vary between individuals within the same population group, as well as between different population groups (Venter, 2021a). As such, assessment of ROS is considered an important factor for the purposes of this study.

ROS is a collective name for reactive oxidative molecules that include superoxide radical ($O_2^{\bullet -}$), singlet oxygen ($^1O_2$), hydrogen peroxide (H_2O_2) and hydroxyl radical ($OH^{\bullet}$) (Mittler, 2017) (top-right diagram in Figure 2.2). Chainy and Sahoo (2020) state that ROS is invariably associated with oxygen-rich metabolism, continuously counteracted by an antioxidant defence comprising reactions spread throughout the body. Additional findings by Chainy and Sahoo (2020) regarding ROS are primarily discussed in the following paragraph.

ROS are normally introduced internally along hormone-regulated bodily pathways, but its presence can also be attributed to external factors such as diet and radiation. ROS-forming pathways can be separated into two groups, namely those with incorporated antioxidant-systems, and those where activation of potent free radicals occur. The former involves melatonin functions, mitochondrial adenosine triphosphate (ATP) formation and oestrogen metabolism, while the formation of insulin, thyroid hormone and/ or corticosteroids and catecholamines metabolism occur within the latter. Imbalances within these pathway groups generally either cause disease or are side effects of a disease. Therefore, it should be noted that ROS could be useful signalling molecules, albeit also being a damaging by-product (Mittler, 2017). ROS signalling is strictly regulated by a variety of proteins in redox mechanisms, transcription factors and DNA (Jumuddin, 2018; Ray *et al.,* 2012).

Redox imbalances attributed to excess ROS along with insufficient capacity to initiate an effective antioxidant response, defined as ‘oxidative stress’, adversely results in oxidative reactions with nearby biological molecules to ultimately overwhelm damage repair systems (Mittler, 2017; Ray *et al.,* 2012). DNA, being one of the affected biological molecules, is damaged by ROS in three different ways, namely DNA strand breakages (removal of a nucleotide base), ROS-produced hydroperoxide lipid-mediated exocyclic DNA-adducts (disrupts binding between DNA strands) and 8-hydroxylation of guanine (Roy & Liehr,1999). The latter is considered a more complex form of DNA damage with a roughly 1.1% chance of mutation by substituting guanine with thymine during either DNA replication misreads or repair of a more frequently occurring 8-hydroxy guanine to adenine mispairing (Cheng *et al*., 1992). Even in the absence of oxidative stress, elevated ROS levels resulting from hypothyroidism and/ or diabetic oxidative environments (where too high amounts of ROS inhibit insulin function) are considered sufficient evidence of an imbalanced pathway (Chainy & Sahoo, 2020). Additionally, increased ROS levels are also correlated with an increase in

catecholamines or amyloid-induced (diseased-formed abnormal cells discarded into organs) apoptosis (Chainy & Sahoo, 2020).

Conversely, ROS also play a beneficial role within an organism, particularly fulfilling a critical signalling function that promotes cell proliferation and continued cell existence in an undamaged state (Ray *et al.,* 2012). In this light, ROS regulate antioxidant activation, stimulate pro-inflammatory pathways and contribute to immunity in pathogen elimination, amongst other functions (Jumuddin, 2018). Although accumulated damage caused by ROS does not directly lead to cell death, elevated ROS levels trigger various physiological pathways that rapidly result in apoptosis of cells as a result of its signalling function (Bassoy *et al* 2021; Mittler, 2017). H_2O_2 forming part of ROS binds to cysteine residues in proteins thereby signalling a regulation response in order to maintain a functional level of enzymes (Mittler, 2017). A small but steady amount of ROS above the minimum threshold ensures cell proliferation and proper immunity activation while also avoiding the toxic effects of large amount of ROS (Mittler, 2017; Bassoy *et al* 2021).

Given the above-mentioned generalized characteristics, sources and functions of three of the four named ROS (excluding the exited form of O_2, namely $^1O_2$, as the focus will rather be on its' reduced form, $O_2^{\bullet -}$) of specific interest to this study are discussed in the following sections.

2.3.1 ROS from Oestrogen Metabolism

Redox cycles found within oestrogen metabolism, comprising a forward reaction catalysed by the *CYP* or peroxidase enzymes reducing O_2 and a reverse reaction catalysed by *CYP* reductase, produce the ROS by-product superoxide anion radical ($O_2^{\bullet -}$) (Cavalieri & Rogan, 2014; Mittler, 2017). Detoxification of $O_2^{\bullet -}$ by superoxide dismutase (SOD) a scavenging antioxidant enzyme, results in the formation of ROS hydrogen peroxide (H_2O_2), whereafter interaction between H_2O_2 and iron (Fe^{+2}), known as the Fenton reaction, leads to the production of hydroxy radicals ($OH^{\bullet}$) (Mittler, 2017).

It must be noted that $O_2^{\bullet -}$ and $OH^{\bullet}$ represent free radicals, while the chemically reactive H_2O_2 could readily undergo an additional conversion into radical ROS (Jumuddin, 2018). However, the catalysing effect of *COMT* enzymes present along phase II methylation pathways reduces the above-mentioned adverse reactions, resulting in the production of the non-carcinogenic methoxylated oestrogen (Samavat & Kurzer, 2015). Methoxy-oestrogen is an inhibitor of carcinogenic activity, for example, cell proliferation and excessive new blood vessel formation (Samavat & Kurzer, 2015). The conjugation-neutralization of glutathione (GSH) metabolite, the most abundant non-protein thiol, is another pathway that counteracts oxidative damage (Jumuddin, 2018; Chainy & Sahoo, 2020). GSH exhibits three antioxidant functions, namely

preventing ROS formation, scavenging already formed ROS and conjugating quinone oestrogens catalysed by GSH-S-transferase (Jumuddin, 2018).

In summary, oxidative damage due to increased ROS production by the accumulation of oestrogen quinones results from either an imbalance between the phase I and II enzymes, the appearance of excessive substrates, and/ or antioxidant depletion (e.g., SOD or GSH) (Jumuddin, 2018).

2.3.2 ROS from Oestrogen-induced Mitochondria

The mitochondrion metabolism induced by oestrogen through E_2 (from multiple sources) and Ca^{+2} cooperation with optional oestrogen receptor alpha (ERα) interaction, including lipoxygenase and membrane embedded electron transport chain (ETC) enzymes, results in the formation of ROS (Felty *et al*., 2005; Zhang, 2022). Although all of the mechanisms involved in oestrogen and mitochondrial interaction have yet to be fully deciphered, Okoh *et al*. (2011) state that one of these involves the binding of E_2 to ERα/ ERß that activates transcription factors (TF) thereby influencing mitochondrial genome transcription. In turn, oxidative-sensitive TF bind to the oestrogen response element (ERE) in genes implicated in cell growth or ETC protein production, causing localized oxidative damage with subsequent genetic mutations and enhanced mitochondria motility (Felty *et al.,* 2005; Okoh *et al*., 2011). A lesser-known mechanism involves the instant production of perinuclear ROS initiated by E_2 that acts as an anchor for mitochondrion via integrin-dependent receptors (rather than ER) thereby initiating signalling that effects the ETC respiratory function (Felty *et al.,* 2005; Okoh *et al*., 2011). E_2-induced mitochondrial ROS act as signal converters to induce physiological functions, for example, insulin secretion (Chainy & Sahoo, 2020; Felty *et al.,* 2005).

2.3.3 ROS and COC Correlations

Whether naturally produced or originating from consumed synthetics (i.e., COC), oestrogen metabolites could have either pro- or antioxidant characteristics depending on which oestrogen metabolite results along the biotransformation pathways (Okoh *et al*., 2011). The oxidative balance depends on the concentrations of the different resultant oestrogen metabolites (Okoh *et al*., 2011).

Most COC (both oestrogen and progestin components) are after consumption swiftly transported to hepatic cells to enter mitochondria in order to thereafter bind to ER within (Fuller *et al*., 2022; Okoh *et al*., 2011). During studies by several authors (e.g., Cauci *et al*., 2016; Finco *et al*., 2011; Fuller *et al*., 2022) ROS were predominantly determined by the measurement of mitochondrial H_2O_2. Increased levels of measured H_2O_2 were positively

correlated with disruption of mitochondria homeostasis (Fuller *et al*., 2022). However, the precise mechanism through which this disruption occurs, being either direct consumption of the antioxidant defence by COC metabolites that inhibit H_2O_2 counteraction, or indirect depletion thereof by COC metabolites through the formation of H_2O_2, still needs to be defined (Cauci *et al*., 2016).

The studies regarding COC (of unspecified brand) metabolism discussed in the previous paragraphs revealed an increase in H_2O_2 leading to oxidative stress environments, increased lipid peroxidation and diminished antioxidant sources in the form of GSH, *SOD* or quinone reductase (*NQO*) (Cauci *et al*., 2016; Finco *et al*., 2011; Fuller *et al*., 2022). Finco *et al*. (2011) noted that the stress environment could not be countered by the use of antioxidant supplements, although supplements containing catechin could be beneficial to a degree.

2.4 Oestrogen and ROS in Breast Cancer

Breast cancer (BCa) can be caused by numerous different genetic (polymorphisms), metabolic (carcinogenic oestrogen metabolites and ROS), and lifestyle/environmental factors (e.g., epigenetics, diet, xenobiotics, etc.) (Cavalieri & Rogan, 2014). As noted by Roy and Liehr (1999), oestrogen could damage DNA and cause cancer in three different ways, namely oestrogen-DNA adducts, oestrogen-induced DNA modification (through receptor binding) and oestrogen-generated ROS.

2.4.1 Oestrogen-Induced Cancers

The disparities in oestrogen metabolism pathways within BCa cells highlighted metabolic markers where if these markers lean towards similar imbalances within healthy (no detectable cancer) premenopausal women, the women are considered at high risk for BCa (Ziegler *et al.,* 2015). Oestrogens considered carcinogenic in most populations will be discussed in more detail in the following paragraphs, including oestrogen precursors (testosterone and androstenedione), E_2-containing metabolites, hydroxylated oestrogens, oestrogen quinones and oestrogen-DNA adducts (Cavalieri & Rogan, 2016; Männistö & Kaakkola, 1999; Moon *et al*., 2006; Sak, 2017; Samavat & Kurzer, 2015).

Saini *et al*. (2021), as well as Samavat and Kurzer (2015), state that women who did not consume any exogenous oestrogen but still exhibit increased levels of oestrogen precursors (such as testosterone, androstenedione) or oestrogen hormones/conjugates (such as E_1, E_2, or E_1-S typically storing excess oestrogen) demonstrated an increased risk for BCa. Women

with BCa and increased testosterone levels but no particular increase in other hormone metabolites were found to associate with ER- and progesterone receptor- (PR) positive and invasive cancers (Eliassen, 2006).

E_2-derivative metabolites, such as 2-hydroxy-E_2, are more reactive than the E_1 and E_3 counterparts (Samavat & Kurzer, 2015). In turn, 4-hydroxylated-E_2 is considered the most reactive metabolite constituting approximately 5% of circulating oestrogens that bind covalently to ER for a longer period and more effectively when compared to the effects caused by E_2, even after the first phase of biotransformation (Männistö & Kaakkola, 1999; Dumas & Diorio, 2011; Sepkovic & Badlow, 2009). Increased levels of both mentioned metabolites are considered biomarkers for the presence of malignant BCa (Samavat & Kurzer, 2015).

Other carcinogen metabolites of interest are the 16-hydroxylated oestrogens (especially the abundant 16-hydroxylated-E_1) with the same ER-binding capabilities as the parent molecules that could trigger cancer formation by promoting unscheduled DNA biosynthesis (Samavat & Kurzer, 2015; Männistö & Kaakkola, 1999). Unscheduled DNA biosynthesis is the formation of nucleotides outside of the usual mitoses S-phase where 16-hydroxylated oestrogens bind to ER in an attempt to repair DNA damage through nucleotide excision repair (NER) thereby producing an excess of nucleotides (Taioli *et al.,* 1995). Consequently, 16-hydroxylated-E_1 has regularly been associated to BCa and other oestrogen-related cancers (Sepkovic & Badlow, 2009).

Catechol oestrogen-3,4-quinone is a potent carcinogen resulting from oestrogen metabolism that binds to DNA in 1,4 Michael addition formation, rather than to the more frequent 1,6 Michael addition formation, producing unstable oestrogen-DNA adducts (Figure 2.3) (Moon *et al*., 2006; Sak, 2017). These DNA-adducts can yield life-threatening DNA mutations (i.e., changes in enzyme activity or expression) that could be the first critical step in the formation of more frequently occurring BCa, thyroid and/ or ovarian cancer among women, prostate cancer among men, or non-Hodgkin's lymphoma in both (Cavalieri & Rogan, 2016; CANSA, 2017). This is supported by high ratios of 4-E_2-1-N3Ade and 4-E_2-1-N7Gua measured for women either already suffering from BCa or exhibiting a high BCa risk in comparison to controls comprising women of relatively similar age and BMI (Pruthi *et al.*, 2012; Samavat & Kurzer, 2015). However, 2-Methoxy-E_2, one of the few protective metabolites present within cancerous tissue, is found to inhibit cell proliferation when measured in large quantities (Männistö & Kaakkola, 1999).

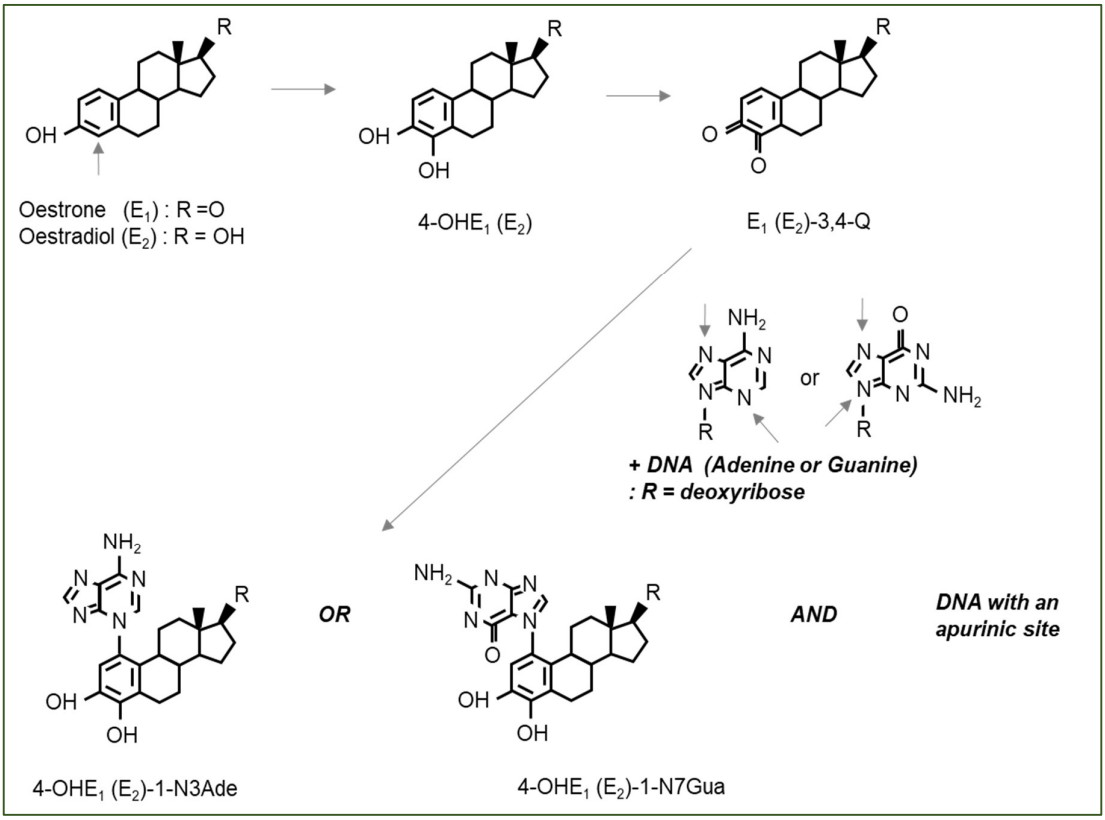


Figure 2.3 Structural presentation of the interaction between the catechol oestrogen quinones and DNA

Legend

OH – Hydoxy, Q-quinone, Ade – Adenine, Gua – Guanine

Adapted from Cavalieri & Rogan (2016)

Users COC have higher levels of oestrogen hormones due to the consumption of synthetic oestrogen and progestin that is similar to that measured during pregnancy, with both of these hormonal profiles correlating with an increased BCa risk (Gaudet, 2005; Fitzpatrick *et al*., 2023). Gierisch *et al*. (2013) state that the long-lasting effect of COC is associated with the aetiology of BCa, although conversely decreased the risk of endometrium cancer. However, Key *et al*. (2001) also mention that individuals still have a 25% increased risk for BCa even a decade after discontinuing use of COC when compared to never users, while users regularly using COC for a period of at least a decade have a 35% increased risk for BCa. It is important to note that some COC associations are not just limited to BCa but extend to all cancers

originating from changes in liver oestrogen metabolism (Key *et al*., 2001). Key *et al*. (2001) also state that different synthesized oestrogens within COC all hold roughly the same cancer risk since all contribute to the same effect. This is further supported by a study done by the UK Clinical Practice Research Datalink (CPRD) that there is a 20-30% increase in BCa for women using EE and DRSP containing COCs or progestin-only contraceptives (Fitzpatrick *et al*., 2023). Even though the use DRSP decrease cell proliferation and therefore decrease BCa risk, contradictory certain subtypes of cancer display an increased cell invasion in the presence of DRSP, increasing the BCa risk displaying a net increase in risk (Van der Meer 2017; Eksteen, 2019).

2.4.2 ROS-induced Cancers

It is known that ROS can either directly damage DNA, protein and lipids with the damaged macromolecules associated with carcinogenesis and aging factors, amongst others, or indirectly signal transcription factors in promoting pro-carcinogenic gene activation (Ray *et al.,* 2012). The production and accumulation of ROS generate an oxidative stress environment that could sustain irregular regulation of proliferation and inhibit apoptosis within cancerous growths (Jumuddin, 2018; Ray *et al.,* 2012).

As hydrogen peroxide (H_2O_2) within the liver is scavenged by reduced glutathione (GSH), resulting in the formation of oxidized glutathione (GSSG) and hydroxyl radicals (OH•), diminished GSH levels indicate a state of oxidative stress (Yuan & Kaplowitz, 2009). OH• reacts with lipids to form hydroperoxides considered relativity unchecked cofactors through *CYP* catalysed reactions that downstream up-regulate carcinogenic oestrogen quinones biosynthesis (Cavalieri & Rogan, 2014).

Mitochondria respiratory chain activity promoted by oestrogen hormones constitutes a major source of ROS within BCa, as well as prostate cancer (Felty *et al.,* 2005). Within cancer cells high levels of ROS lead to mitochondrial dysfunction due to damage to mitochondria DNA, respiratory system, and polarized membranes (Jumuddin, 2018). ROS contribute to signal transduction that activates transcription factors (TF) involved in G1-phase expression genes and could result in accelerated mitosis periods (Felty *et al.,* 2005). Some of the ROS signalling pathways assist cancer cells that require more nutrients to adhere to other cells to form a cohesive mass (Felty *et al.,* 2005)

.

2.5 Related Genetics

Although Cavalieri and Rogan (2014) stated that various factors can lead to the formation of cancer, this study mainly focusses on genetic influence while minimizing the effects of other factors. Similar to ROS levels, breast cancer (BCa) formation and presentation differ between populations and also between individuals of the same population, making it a highly heterogenous disease that complicates diagnosis and treatment (Huo *et al*., 2009). Characteristics of BCa applicable to both African and Caucasian populations are discussed in the following section, whereafter differences in this regard will follow.

2.5.1 Biotransformation Genes involved in Cancers

Genetic polymorphisms influencing the activity and expression of genes encoding oestrogen metabolism enzymes could subsequently contribute to or counteract cancer risk (Samavat & Kurzer, 2015). This is corroborated by the results of several studies on phase I pro-carcinogen activators, namely *CYP1A1*, *CYP1A2*, *CYP1B1*, *CYP3A4* genes and the primary phase II inactivator *COMT* gene (Bai *et al.,* 2017; Long *et al.,* 2007; Samavat & Kurzer, 2015).

Other proteins, such as receptors, transporters, and repair enzymes, are also involved in cancerous aetiology and function (Liu & Lu, 2020; Samavat & Kurzer, 2015; Savage *et al*., 2014). Steroidogenic acute regulatory protein (StAR) regularly transfers cholesterol into the mitochondria for oestrogen biosynthesis, membrane maintenance and bile biosynthesis within ovarian cancer cells (Samavat & Kurzer, 2015). Polymorphisms of the *BRCA* gene encode a repair enzyme that is generally inefficient in the repair of double strand breaks (DSB) by means of homologous repair (HR) within DNA, resulting in significant genome instability while promoting cancer formation (Liu & Lu, 2020; Savage *et al*., 2014).

Polymorphisms that contribute to increased activity or expression of pro-carcinogenic enzymes (especially *CYP1B1)* and/or decreased activity or expression of protective enzymes (such as DNA repair enzymes) are associated with cancer formation in premenopausal women (Cavalieri & Rogan, 2014; Jumuddin, 2018; Moon *et al*., 2006; Sak, 2017; Savage *et al*., 2014). Genes that encode enzymes promoting cancer formation are usually expressed in large quantities within cancerous tissue (Beuten *et al.,* 2008).

In this light, the *CYP1* family of genes are considered to produce about two thirds of toxic metabolites considered to exhibit possible cancer-forming characteristics (Alsubait *et al.,* 2020). *CYP1A1* and *CYP1A2* have multiple polymorphisms that are present in most populations groups and are associated with BCa (Dumas & Diorio, 2011; Long *et al.,* 2017). While most *CYP* enzymes are present in large quantities within cancerous tissues, reduced levels of *CYP1A1* enzymes occur after cancer formation (Dumas & Diorio, 2011). It is

important to note that the activity of the *CYP1A2* enzyme is influenced by numerous consumed and environmentally-absorbed metabolites and consequently could demonstrate a 40-fold variance (Faber *et al*., 2005; Liska *et al*., 2006). Although most polymorphisms in *CYP1A2* genes up-regulate expression, polymorphisms located within the promoter region thereof are considered to decrease its activity (Bu *et al.,* 2014; Faber *et al*., 2005)

The polymorphism that up-regulate the *CYP1B1* enzymes have been associated with multiple types of BCa, as well as brain, colon, ovarian and prostate cancers, but conversely with a decrease in endometrium cancer risk (Alsubait *et al.,* 2020; Ashton *et al.,* 2010; Zahid *et al.,* 2018). However, *CYP1B1*-catalysed metabolism occasionally plays a beneficial role by detoxifying specific xenobiotics, while some polymorphisms down-regulate *CYP1B1* enzyme expression demonstrating a protective cancer survival effect (Ashton *et al.,* 2009; Long *et al.,* 2007).

Polymorphisms in *CYP3A4* correlate with changes in oestrogen metabolism but only to a degree (Katzung, 2017). Some of these *CYP3A4* polymorphisms have been associated with oestrogen-induced cancer, notably BCa, although the precise effect of the polymorphisms on enzyme activity or expression and oestrogen metabolism in the absence of outside factors is yet to be discovered (Bai *et al.,* 2017).

Finally, *COMT* is a complex major phase II enzyme, the characteristics of which is best described by Sak (2017). Polymorphisms within *COMT* have a significant impact on enzyme activity on both the membrane-bound and soluble form. These polymorphisms and high levels of E_2 metabolites down-regulate the primary phase II methylation enzyme, and increase the risk of lethal subtypes of breast, endometrium, and ovarian cancer in women and prostate cancer in men. Decreased *COMT* activity also leads to the formation of new blood vessels while the lack of methoxy-oestrogen hinders apoptosis, thus promoting cancer growth. *COMT* is reported to have complementary interactions with other genes while also reacting to environmental stimuli (such as xenobiotics) but further study is required to facilitate better understanding of links between *COMT* enzymes and cancer. The heterogenic nature of *COMT* enzyme activity has been measured in multiple populations, indicating that polymorphism do not constitute a singular biomarker for cancer formation. Even after extensive research, assessment of the association between *COMT* enzyme activity and cancer still rarely yields reproducible results.

The multiple anti-carcinogenic genes involved in oestrogen metabolism that could possibly interact with the above-mentioned genes are summarized in this section. A decrease in the activity of the enzyme encoded by the *COMT* gene along the most favoured phase II pathway results in saturation and measurable increase in activity of the competitor quinone-forming enzymes (e.g., *CYP* and peroxidase) (Cavalieri & Rogan, 2014). In response to the higher levels of oestrogen quinones, an anti-carcinogenic protective enzyme known as NAD(P)H

quinone oxidoreductase 1 encoded by the *NQO* gene subsequently reverses quinone formation reaction while producing CE allowing favoured phase II oestrogen metabolism (Cavalieri & Rogan, 2014; Jumuddin, 2018). Another set of anti-carcinogenic genes causes the restoration of oxidated lipids catalysed by glutathione peroxidase, encoded by the *GPx* gene, while transforming the antioxidant, reduced glutathione (GSH) into the oxidized glutathione (GSSG) within the liver (Yuan & Kaplowitz, 2009).

2.5.2 Differences among Population Groups

All five of the genes considered of interest for the purposes of this study (i.e., *CYP1A1*, *CYP1A2, CYP1B1, CYP3A4* and *COMT*) have been confirmed to influence phenotypical differences and presentation between ethnic groups, for instance African women that could exhibit slightly different oestrogen levels (Faber *et al*., 2005; Li *et al.,* 2014; Reding *et al*., 2012). However, there are enough similarities between the genetics of African American and that of indigenous African women to be able to link and collate the results of various association analyses studies (Reding *et al*., 2012).

Although Reding *et al*. (2012) state that the lifetime breast cancer risk is higher in Caucasian women, African women were found to exhibit more malignant and therapy-insensitive cancers. That said, approximately 35% of breast cancer within African women is basal-like (i.e., malignant) compared to only 16% in Caucasian women, albeit equally aggressive in both populations (Ademuyiwa *et al*., 2017). This malignant cancer type is predominantly a triple negative (TN) subtype that does not express ER, progesterone receptor (PR) or human epidermal growth factor receptor 2 (HER2) required to facilitate treatment via hormone-therapy (Ademuyiwa *et al*., 2017; CANSA, 2017).

Breast cancer is considered the most relevant cancer for all South African women and prostate cancer the most relevant in men, where non-Hodgkin's lymphoma is included in the top-5 cancers for both genders according to CANSA (2017). Some *CYP1A1, CYP1B1* and *CYP3A4* polymorphisms are associated with BCa in Caucasian and African populations, while some *CYP1A2* polymorphism associated with an ER-positive BCa were observed only in African American women, demonstrating differences in genetic arrangements found within BCa cases (Ademuyiwa *et al*., 2017; Beuten *et al.,* 2008; Quan *et al*., 2014; Samavat & Kurzer, 2015; Taioli *et al*., 1995; Werk & Cascorbi, 2014; Zahid *et al.,* 2018). Additionally, BCa is known to form in African American women at a relatively young age in contrast to its occurrence in older Caucasian women (Amend *et al.,* 2006). Differences in genetic polymorphism frequencies between populations have been associated with a higher risk of prostate cancer formation in the Caucasian population (Beuten *et al.,* 2008). However, the *CYP3A4* gene exhibits an increase in polymorphism frequencies within the African American populations when

compared to that within the Caucasian population (Li *et al.,* 2014). While differences in oestrogen metabolites are typically associated with BCa and *COMT* polymorphisms, some studies have shown that BCa associate with COMT polymorphisms in both populations while other studies found no such association (Janacova, 2015; Reding *et al*., 2012). Additionally, contrary to African American men in whom *COMT* polymorphisms that resulted in low enzyme activity strongly associated with increased prostate cancer risk, Caucasian men with same polymorphisms did not associate with any risk of prostate cancer (Reding *et al*., 2012; Sak, 2017).

In order to describe the precise effect of a specific polymorphism on human health, more in-depth studies involving different ethnicities should be conducted while keeping external factors and genetic interaction in mind (Bai *et al.,* 2017).

2.6 Association Analysis Relevance

Nucleotides consisting of a phosphate group, pentose sugar and one of four different nitrogen bases form the building blocks of DNA (Brookes, 1999). A single nucleotide polymorphism (SNP) defines a specific location on the genome where two or more different bases (known in this context as alleles) could naturally occur, and where the minor allele frequency (MAF), being the second-most abundant allele, comprises at least 1% within any population (Brody, 2016). Statistical analyses of the association of these SNPs with disease or specific metabolites could lead to better understanding of phenotypical variety (Frazer *et al.*, 2009).

2.6.1 Basic Principles

Every individual possesses two copies of each chromosome and by extension two of each gene or specific nitrogen base, one inherited from each biological parent (Brody, 2016). SNPs are mostly bi-allelic, which implies that the two inherited bases could combine two alleles in three different combinations (Figure 2.4) (Brookes, 1999; Frazer *et al.*, 2009). Other SNPs can be tri-allelic where two inherited bases could combine three alleles in six different combinations, as demonstrated by the different blood types with dominant and recessive characteristics (Männistö & Kaakkola, 1999).

A SNP is occasionally located within a coding region (exon) that could cause ammino acid substitution, and as such influence enzyme activity (Brody, 2016). Alternatively, the SNP could be located within a regulatory region (intron), or a region between genes, resulting in a change in enzyme expression rates (Brody, 2016). The third combination in Figure 2.4 provides an

example of a SNP comprising two inherited copies of a minor allele that exhibits slightly slower activity occurs within a coding region thereby potentially limiting the metabolism of specific substrates within an individual (Liska *et al*., 2006).

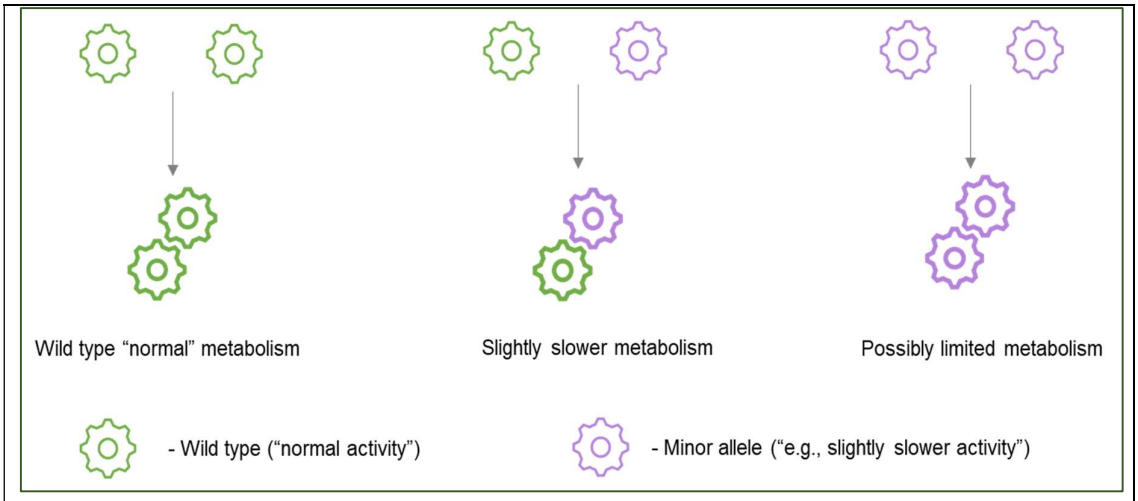


Figure 2.4 Illustration of bi-allelic inheritance combinations

Adapted from: Liska *et al*., 2006 and Brody (2016)

Changes in DNA nucleotides that occur in less than 1% of a population are known as a mutation being either life-threatening or newly formed (Frazer *et al.*, 2009). Of the approximately 11 million SNPs that have been recorded, a total of seven million have a MAF greater than 5% (Frazer *et al*., 2009).

As SNPs are known to contribute to phenotype diversity, further study of SNPs could reveal more regarding pathway function, gene interactions and the aetiology of diseases on a molecular level (Frazer *et al*., 2009).

2.6.2 SNP to Metabolite Association

Although numerous studies associate SNPs with either a disease or disorder, this study rather focusses on associations between SNPs and metabolites in order to improve understanding of the role that SNPs play in pathways within biological mechanisms (Frazer *et al*., 2009).

Association analysis entails the processing of hundreds to thousands of SNPs to produce only a few significant SNP-metabolite associations (Nam *et al.,* 2010). However, use of association analysis has been found to be the most effective when the number of SNPs are limited to pre-defined genes potentially linked to complex traits to avoid discarding significant SNPs during statistical adjustment (Nam *et al.,* 2010; Zhang *et al.,* 2002).

However, use of statistical association to analyse the correlation between a SNP and a functional difference in a complex phenotype (e.g., slower metaboliser enzyme) overlooks the role that other factors play in the change in enzyme expression or activity (Frazer *et al.*, 2009). Additionally, an association found in one population cannot be assigned to other populations, while not all of the SNPs or other factors that contribute to changes in phenotype have been discovered yet (Frazer *et al.*, 2009). Another limitation of the use of association analysis is a lack of sufficient statistical power to detect rare mutations involved in serious diseases, the effects of which are omitted during these calculations (Frazer *et al.*, 2009).

Lastly, results of association studies within the same population have been reported to be inconsistent due to several factors, as summarized in this section (Sak, 2017). These factors include insufficient sample size, group selection (external factors control), metabolite availability within cells, genetic interaction with the environment, and MAF that differs between populations. Additionally, SNP association should not be oversimplified as the expression of genes (e.g., *COMT*) requires several environmental and lifestyle stimuli regardless of SNP influences. Structural changes within DNA could also alter gene expression (e.g., hypermethylation of the gene promoter region) thereby inhibiting gene expression. Finally, it must be noted that a single SNP is rarely the only contributor to a highly regulated oestrogen biotransformation enzyme.

Gabriel *et al*. (2002) state that association of haplotypes (SNPs inherited together due to close proximity and considered less prone to recombine) with metabolites is an alternative use of association analysis. Furthermore, Zahid *et al.* (2018) are of the opinion that association analysis using a polygenic approach based on the interaction between genes could further understanding of biological mechanisms.

In closing, it is important to note that this study will refer to specific allele association rather than assuming that SNP or polymorphism association refers to the minor allele effect as is more common in practice (Brody, 2016).

CHAPTER 3 – METHODS AND MATERIALS

Literature regarding specific single nucleotide polymorphism (SNP)-metabolite association studies within African populations is scarce, indicating the necessity to gain knowledge about these specific SNP-metabolite associations. Most studies have been done on Caucasian populations and have shown that within a population blood and urine metabolite levels can vary tremendously – even after individuals have consumed the same dose and type of medications (Katzung, 2017). These differences may be caused by SNPs in the biotransformation enzymes. However, SNP-metabolite associations studies that look into this are lacking – even in the Caucasian population which is in contrast with highly studied drug metabolism. This study makes a contribution to partially fill this lack of knowledge, and this chapter describes the reproducible steps taken during sample preparation and quality control, sample analysis, data processing and quality control, and association analysis of specific SNPs with metabolites from the oestrogen biotransformation pathway and markers of oxidative stress in healthy South-African women of African and Caucasian descent.

3.1 Experimental Design Flowchart

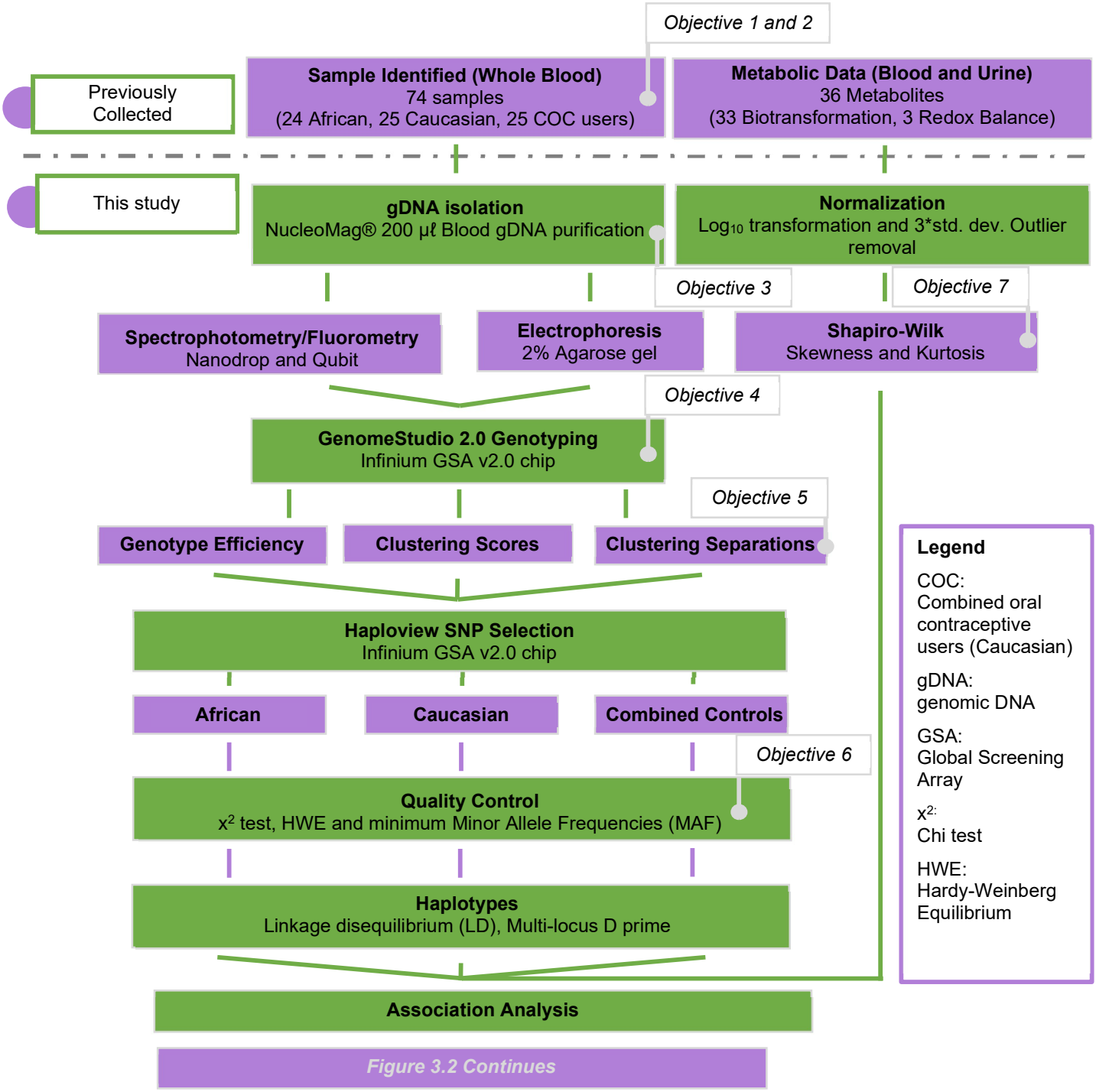


Figure 3.1 Sample and Data Processing Flowchart

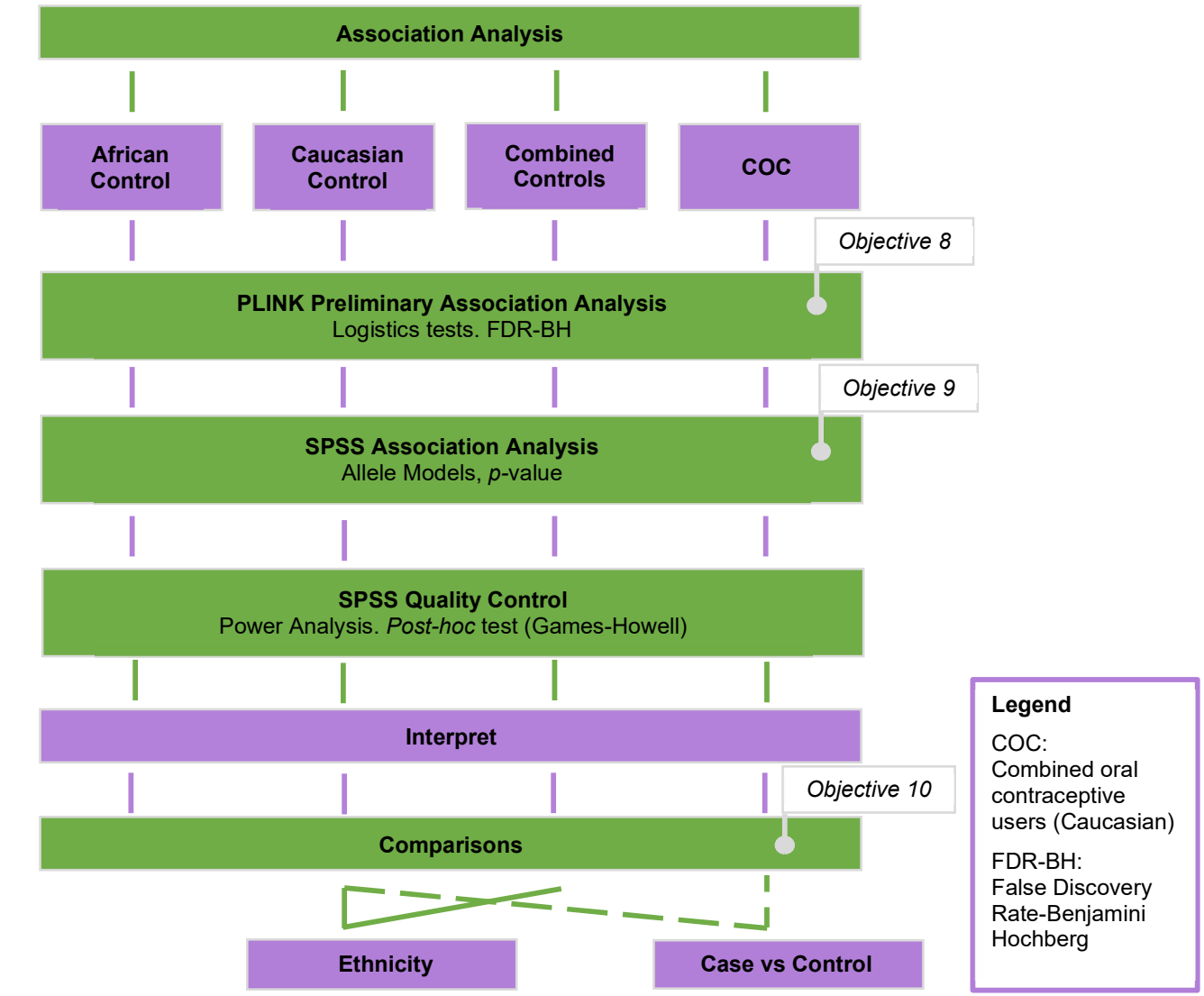


Figure 3.2 Association Analysis Flowchart

3.2 Samples Previously Collected and Study Groups Identified

Anticoagulant (EDTA)-treated whole blood samples of 74 women were retrieved from -80 °C storage. These 74 were the only samples that could be specifically selected from hundreds of volunteers in order to minimize external factors. The samples were collected from participants of the Oestrogen, Biotransformation and Oxidative Stress Status (eBOSS) study (ethics approval number NWU-00344-16-A1). Informed consent was obtained from all participants to collect and store blood samples for genetic analysis. All the participants were of the ages 18-35 years, with a BMI below 30 and most were in the luteal menstrual phase during sampling along with many more criteria listed in Venter *et al.,* (2021b:3). There was a total of 74 participants and consisted of 24 Africans and 50 Caucasians. The Caucasian group consisted of two subgroups of 25 combined oral contraceptive (COC)-users and 25 Controls. The COCs used by the women contained 30 mg drospirenone and either 20 or 30 µg ethinyl oestradiol. Due to very low user numbers for these specific COCs in African women, an African COC user group could not be included in the eBOSS study. This genetic analysis study was approved by the North-West University Health Research Ethics Committee (NWU-00417-20-A1).

Table 3.1 Arranged Participation Groups from previously collected eBOSS samples

Group Number	Group Name	Number of participants
Group 1	Caucasian Control	25
Group 2	African Control	24
Group 3	African and Caucasian Controls (Combined Controls)	49
Group 4	(Caucasian) Combined oral contraceptive (COC) users	25
Total: 50 Caucasians; 24 Africans		

This information was used to achieve objective 1.

3.3 Identified Genes

Within the oestrogen biosynthesis, transportation and metabolism pathways are hundreds of catalysing enzymes encoded by hundreds of genes (Liska *et al*., 2006). Oestrogen undergoes phase I and phase II biotransformation within the liver. The catechol oestrogens (CE) produced during phase I are ideally methylated and excreted. However excess formation of

CE metabolites can result in their conversion into oestrogen quinones (a reaction that is coupled to the formation of ROS). Reduction of quinones reverses the reaction back to form CE results in the redox cycling reaction (which occurs between the two phases) that results in the increased ROS formation (Li *et al*., 2017; Liska *et al.,* 2006). These ROS could damage surrounding protein, lipids and DNA (Ray *et al.,* 2012).

The repair of DNA could cause mutations which leads to trigger the formation of breast or other cancers (Cavalieri & Rogan, 2014). Although there are numerous *CYP* enzymes involved in phase I oestrogen biotransformation and more knowledge was gained of different dominant enzymes in oestrogen metabolism between different populations, the *CYP1A1,1A2, 1B1* and *3A4* genes were initially selected as the primary phase I enzymes (Alsubait *et al.,* 2020; Katzung, 2017 and Liska *et al.,* 2006). The most effective and dominant phase II enzyme, *COMT*, was also included (Sak, 2017). These selected genes corresponded to179 SNPs from the Infinium Global Screening Array v2.0 (GSA v 2.0) Beadchip.

During the eBOSS study project (NWU-00344-16-A1), 36 metabolites relating to oestrogen metabolism and markers of oxidative stress were analysed. The results obtained have not yet been published yet but were discussed in person with the researcher (Venter, 2021a). There were significant differences between individuals of the same ethnicity and between ethnic groups, even though the BMI and age were similar as will be confirmed in this study's results which is a similar to that described by Key (2011) and Pruthi *et al.* (2012) studies. The five selected genes encode enzymes that could mostly directly influence the measured oestrogen metabolites and indirectly influence ROS. This part of the methodology was employed to meet the requirements of objective 2.

3.4 gDNA Isolation

In order to determine the genotype of each individual, firstly the genetic material needed to be isolated. The concentration and quality of the DNA were assessed **spectrophotometrically,** and the integrity of the DNA was examined with **electrophoresis and UV imaging of agarose gel** as requested by the Centre of Proteomics and Genetic Research (CPGR) (Figure 3.3).

SAMPLE REQUIREMENTS	DNA
Infinium Arrays	A260/A280 ratio 1.7 to 2.1 A260/A230 ratio 1.5 – 3.0 500ng High MW DNA free of inhibitors & contaminates. Concentration 50ng/µl

Figure 3.3 gDNA Quality Control Requirements (CPGR, 2022)

3.4.1 Magnetic gDNA isolation

The first step was to isolate genomic DNA (gDNA) from the whole blood samples using the Omega Bio-tek Mag-BIND® Blood kit. This kit used a lysis Buffer and proteinase K (Prot. K) to break down the cell membranes within blood and dissociate proteins from DNA to yield free gDNA pieces within the mixture. The paramagnetic beads (weak, reversible adsorption binding beads) were added to bind to the gDNA. After washing away contaminants and salts, an elution buffer was added to separate the gDNA from the magnetic beads to be extracted as pure gDNA. The steps that were followed for isolation of gDNA are shown below in Figure 3.4.

Step 1 Prot. K preparation – Dissolve 50 mg Prot. K powder in 2.5 mℓ Proteinase Buffer PB

Step 2 Lyse samples – Mix 20 µℓ Prot. K. with 200 µℓ blood and 80 µℓ MBL 1 (Lysis Buffer) Briefly vortex 3-5 times, then shake at room temperature (RT) for 10 min.

Step 3 Bind gDNA – Mix 25 µℓ B-Beads with 300 µℓ MBL 2 (Binding Buffer). Briefly vortex. shake at RT for 5 min. Induce magnet separation for 2 min. Remove the supernatant.

Step 4 Wash gDNA – Add 800 µℓ MBL 3 (Wash Buffer) to gDNA, pipette up and down 15 times, and shake at RT for 5 min. Induce magnet separation for 2 min. Remove the supernatant. **(Repeat once more)**

Step 5 Wash gDNA – Add 800 µℓ freshly made 80 % ethanol to gDNA pipette up and down 15 times, and shake at RT for 5 min. Induce magnet separation for 2 min. Remove the supernatant.

Step 6 Wash gDNA – Leave the tube connected to the magnetic separator during this step. Slowly add 900 µℓ MBL 4 (Wash Buffer). Gently remove supernatant within 45-90 s. Discard supernatant.

Step 7 Elute DNA – Add 50 µℓ MBL 5 (Elution Buffer) at ≈ 70 °C. Shake at RT for 10 min. Induce magnet separation for 2 min. Transfer eluted gDNA to tubes.

Figure 3.4 gDNA isolation protocol using Omega Bio-tek's Mag-BIND® Blood kit Protocol

3.4.2 Spectrophotometric Quality Control

Quality Control measurements were necessary to ensure accurate genotyping. gDNA purity was assessed spectrophotometrically using the Nanodrop One / One^{C} Microvolume UV – Vis Spectrophotometer from ThermoFisher Scientific to detect any protein and other contaminants. The concentration of the isolated gDNA was determined by another instrument, namely the Invitrogen Qubit 2.0 Fluorometer instrument by Life Technologies.

The Nanodrop is quite well described by Koetsier and Cantor (2019). The Nanodrop instrument measures UV absorbance with a monochromatic light detector. All molecules absorb various wavelengths at different strength and in order to identify specific molecule presence we measure the wavelength where the molecules absorb the most. With the application of Beer-Lamberts Law, the absorbance is measured. Salts absorb at 230nm, nucleic acid at 260nm and protein at 280nm, and by measuring the absorbance at these wavelengths, one can determine nucleic acid (double and single strand DNA (dsDNA / ssDNA), and RNA) concentration and purity, as shown in Figure 3.6

The A_{260}/A_{280} and A_{260}/A_{230} ratios measures nucleic acid purity. A_{260}/A_{280} at a range of 1.7-2.1 could indicate possible pure DNA. A high A_{260}/A_{280} ratio could indicate RNA contamination and a low value protein contamination. The A_{260}/A_{230} ratio is quite sensitive and a low ratio could indicate guanidine contamination from the previous Omega Bio-tek isolations steps.

For every measurement, 2 µℓ of (sample) volume was applied to the pedestal. The adhesion capillary properties of the light source emit different wavelengths from above through the 2 µℓ column and the detector measures the absorbance (Figure 3.5). The MBL 5 (Elution Buffer) was used as a blank/comparative reading. The pedestal was wiped clean after each measurement. The isolated gDNA sample measurements were done in triplicate to obtain accurate A_{260}/A_{280} and A_{260}/A_{230} ratios.

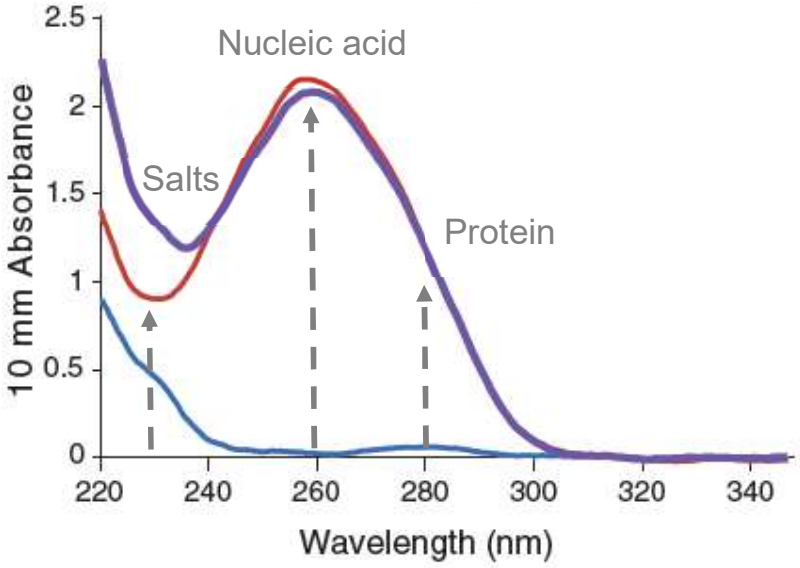


Figure 3.6 Illustration of absorbance maxima for salts, nucleic acids and proteins at specific wavelengths

Adapted from (Koetsier & Cantor, 2019).

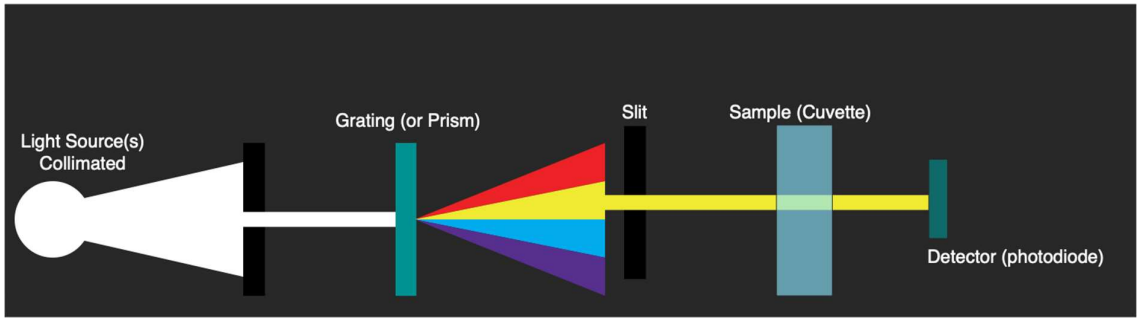


Figure 3.5 UV adsorption measurement (Whiteford, 2021)

The Qubit Fluorometer was used to more accurately determine DNA concentration compared to the Nanodrop concentration measurements and is highly recommended especially for gDNA (Koetsier & Cantor, 2019; Li *et al.,* 2021). How a fluorometer functions and measures is illustrated by Figure. 3.7. UV and visible light is emitted from the light source, the excitation filter removes the visible light and lets the UV light pass through (Lakowicz, 1999). The fluorometers depend on fluorescence of dyes that bind to target molecules which the UV light will excite when in contact (Lakowicz, 1999). During sample preparation, the fluorescent dye binds to the minor groove of dsDNA and changes conformation to emit fluorescence when excited by a light source (Li *et al.,* 2021). Within the excited dye there is the rapid transition of

electrons to a lower energy state which is accompanied by re-emission of absorbed energy, causing fluorescent light to be emitted (Li *et al.,* 2021). The dyes are highly photo-stable with high brightness and is less influenced by other molecule contamination (Li *et al.,* 2021). At a 90° angle another filter allows the fluorescent light through and not the scattered UV light enabling the detector to measure the intensity of the emitted fluorescent light determining the concentration of sample (Lakowicz, 1999).

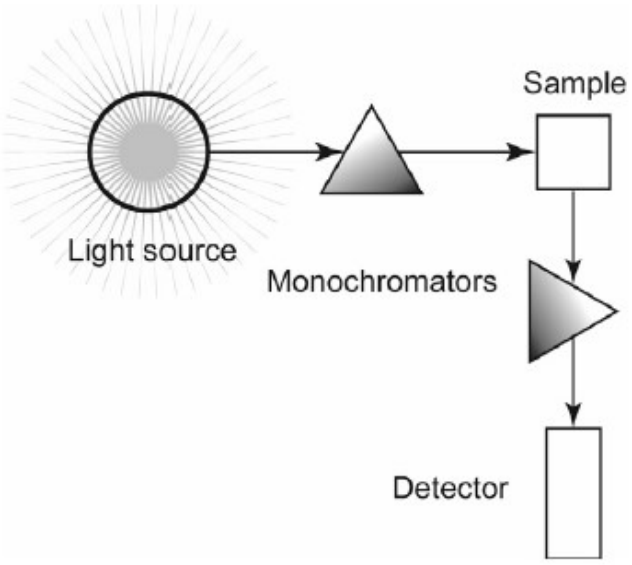


Figure 3.7 Illustration of Qubit fluorescence excitation and detection set-up

All samples were prepared at RT as demonstrated by Figure 3.8 and detailed in the Quick Reference protocol (ThermoFisher Scientific, 2021). Briefly, as noted by Life Technologies (2015), 200 µℓ working solution was made per sample from 1 µℓ Qubit reagent (fluorescent dye and dimethylsulfoxide (DMSO) solvent) and 199 µℓ Qubit buffer (balances pH and provides reaction environment). Two standards solutions were made by adding 190 µℓ working solution and 10 µℓ of the two dsDNA standards provided with the kit (fixed concentration to calibrate the Qubit instrument). Samples were prepared by adding 1 µℓ of gDNA sample and 199 µℓ working solution. All tubes had to contain 200 µℓ of liquid mixtures. After two minutes of mixing, the standard and sample solutions were considered stable and had to be measured within three hours (Life Technologies. 2015).

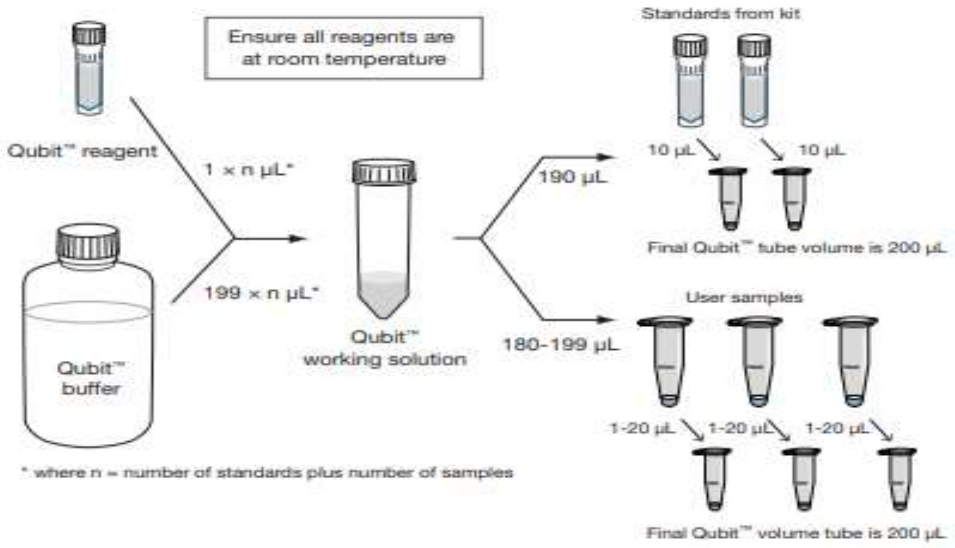


Figure 3.8 Qubit Assay, MAN0017210, Quick Reference Protocol (ThermoFisher Scientific, 2021)

3.4.3 Electrophoresis

Electrophoresis is an effective method for separating varying DNA fragments and varying sizes (Lee *et al*., 2012b) and to determine integrity of the gDNA as well as the presence of RNA by separating the molecules by size and electrical charge (Arslan *et al*., 2021). The DNA has a negatively charged, phosphate backbone that within an electrical current will migrate to the positive anode (Lee *et al*., 2012b). Furthermore, the medium necessary for size-based separation is formed when agarose polymerizes into a gel. During polymerization, the molecules link non-covalently into a mass with specific sized pores depended on the amount of agarose present (Lee *et al*., 2012b). The DNA has a fixed ratio between electrical charge and mass and thus DNA of the same size will uniformly migrate at the same rate (Arslan *et al*., 2021) The distance migrated is inversely dependent of the DNA molecular weight, meaning that smaller DNA fragments will travel farther away from the initial wells (Lee *et al*., 2012b).

There are numerous factors that determine how DNA migrates through agarose gel, with the most well known to be DNA size, agarose concentration (the pore size), DNA conformation (circular, linear or supercoiled), the voltage applied to the electrical current and the type of electrophoresis buffer used (Lee *et al*., 2012b).

Most electrophoresis protocols are based on what Johansson (1972) described and is summarised this paragraph. First, the 2% Agarose gel was prepared by dissolving 1.00 g of agarose powder in 50 mℓ of Tris/Borate/EDTA (TBE) buffer (or 4.00 g of agarose in 200 mℓ buffer). Ethidium Bromide (EtBr; 1 µℓ) was added for every 50 mℓ of TBE buffer. One 50 mℓ agarose gel with a single comb and three 200 mℓ agarose gels with two combs in parallel, were made. The first lane (from the left) was loaded with 2 µℓ of a 17 Kb ladder or a 100 bp ladder (SM0241) as the positive controls. Elution buffer (5 µℓ) was used as the negative control in the second well of the top lanes of each gel. The gDNA samples (1 µℓ) were combined with 4 µℓ 6x loading dye and loaded onto the gel in numerical order. The electrophoresis chambers were filled with TBE buffer until the agarose gel was fully covered and then connected to a Bio Rad power source through which a voltage of 40V was applied for half an hour. Thereafter, the voltage was increased to 70V for two hours, then the gel was removed for imaging.

The gel was imagined using the Bio Rad ChemiDoc™ Imaging System and works similar to the Qubit fluorescence as discussed in 3.3.2. The EtBr that the agarose gel is stained with binds to the migrated DNA grooves and changes shape to emit fluorescence when excited by a UV light (Kaabouch & Schultz, 2007). A camera within the machine adjusts and take a picture of the fluorescent bands on a gel and displays it on a screen to be further interpreted (Kaabouch & Schultz, 2007).

3.5 Genotyping and Quality Control

The gDNA samples were sent to CPGR in Cape Town for further analysis using the Infinium Global Screening Array v2.0 (GSA v 2.0) Beadchip and Illumina iScan system to detect the genotype of hundred thousands of multiple targeted SNPs. The Beadchip is cost effective and can detect population scale variations (Illumina, 2018). The GSA v 2.0 Beadchip was selected since the chip contained multiple liver biotransformation, oestrogen–related and oxidative stress-related SNPs (Illumina, 2018) listed in human genome GRCh37. Also, a lot of SNPs that are disease markers such as 2 260 cancer-related SNPs from numerous databases and pharmacogenetics influence (adsorption, distribution, metabolism and excretion) are also included in this Beadchip (Illumina, 2018). This is a suitable Beadchip for risk profile analysis and population variation determination.

The Infinium GSA v 2.0 Beadchip loading and imaging was done by CPGR in Cape Town, although the workings of the chip will be explained in some detail from the sources Adler *et al.* (2013), Illumina (2018) and Mills (2023). Figure 3.9 and 3.10 illustrates the workflow steps of the Beadchip.

Step 1 Amplification – Add the amplification reagent (contains multiple enzymes and nucleotides) to the samples and place within a 37°C chamber for 20-24 hours. The pure gDNA sample is denatured into ssDNA and therefore anneals to a complementary oligo-nucleotide string. The DNA is amplified by enzymes within the amplification reagent a thousand-fold in this isothermal reaction.

Step 2 Fragmentation – Add the fragmentation reagent at 37°C and leave in a head block for one hour. The now amplified DNA will be cleaved into 300-600 bp fragments. These fragments are the optimal length for Infinium Beadchip hybridization.

Step 3 Precipitation – Add the precipitation reagent (binds to fragmented DNA) at room temperature (RT) and incubate the mixture for 5 min. in the 37°C heat block. Add isopropanol (salt that negates DNA negative charge) and wait for 30 min. in a 4°C environment. Centrifuge mixture for 30 min in 4°C at 3000×g. Carefully remove the supernatant by decanting and take care not to contaminate or dislodge the precipitated DNA pellet.

Step 4 Resuspend – Add the fresh Resuspend reagent to the pellet in the 48°C oven for one hour and mix well before and after incubation.

Step 5 Hybridization – Prepare with care the Beadchips. There are many instructions and sensitivities to adhere to. Dispense the DNA fragments of a sample into the specified well in the chip. The number of samples a chip can contain depends on the size of the chip. Incubate in the 95°C oven for 16-24hours. The Beadchip has millions of tiny spherical beads rooted onto the bead surface. Each bead has numerous oligo-nucleotides probes that are related to a specific SNP per bead. Two types of beads/probes used in this Beadchip with Infinium II as the most common type. Within the 16-24 hours the DNA fragments anneal to a complementary probe right to the adjacent base that defines the SNP. The probe length, the temperature and stringent buffer ensure that the specificity of the bindings are as accurate as possible. After incubation, immediately submerge the Beadchips in washing buffer. Place the glass inclined edge face-down on the bead to enclose the bead and create a well to add staining liquid to.

Step 6 **“**Xstain” (Extension and Staining); Extension – Assemble a flow-through chamber and heat to 44°C. Added in multiple steps are single nucleotide bases known as chain terminating dideoxy nucleotides (ddNTP). The smaller pyrimidine A and T ddNTP binds to a dinitrophenyl (DNP) and the bigger G and C purines ddNTP bind to biotin. The fragmented DNA that binds to the complementary oligonucleotide is extended by a single nucleotide base by polymerase enzymes. Each sample has two nucleotide

bases per SNP. The samples that are homozygous (the exact same base twice) will have identical ddNTP labels, where heterozygous will have different ddNTP bound to the oligonucleotide on the same bead for a specific SNP.

Staining – Apply specific fluorescent-labeled probe to the well of the beadchip in the flow-through chamber. The green (streptavidin) fluorescence binds to the biotin labelled ddNTP and the red (antibody-DNP) fluorescence binds to DNP. There are consecutive rounds and multiple fluorescence probes bind to a single- ddNTP to amplify the fluorescence signal. The most common Infinium II has a single type of beadchip/probe that hybridizes to both alleles of the SNP such as A/G, A/C, T/G and T/C. In the cases where a SNP has two alleles that will display the same fluorescent dye (e.g., A and T) the less common Infinium I method uses two beadschip/probe types to bind each to one specific allele's adjacent nucleotide and measures signal instead of fluorescent colour. This Infinium I uses more space on the beadchip than the Infinium II type does.

Step 7 Imagining – Green and red lasers excite the fluorescence probes, the iScan overlaps the green and red intensities images and will display a yellow colouring for heterozygous SNP samples as shown below. The Illumina iScan system will store intensity files of the green and red fluorescence from the beadchip and a program such as GenomeStudio will convert and compare the fluorescence to the positions on the array SNP manifest for each locus provided by the company in order to make genotype calls for each SNP per sample.

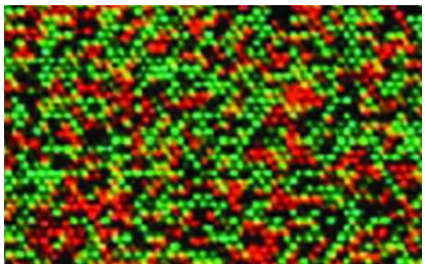


Figure 3.9 Genotyping Preparation and Calls from the Illumina (2018) ; Adler *et al.* (2013) Protocols and Mills (2023)

The genotyping data was visualized, and Quality Control (QC) done using Genome Studio 2.0 Software. The first QC parameter that was considered, was the Call Frequency. The Call Frequency is a percentage measurement of the genotyping efficiency (i.e. for how many of the SNPs on the array a genotype could be accurately assigned) and should be above 95% and preferably 98 – 99%. The other QC markers were GenCall score (GC score) which determines if a sample is close enough to a genotype cluster to be assigned a genotype (Oosting & Oosting, 2014). The GenTrain score (GT score) determines whether the two homozygous groups and the one heterozygous group have properly clustered. Lastly, the Cluster separation score measures if the clusters were separated well enough to ensure that the heterozygous cluster was separately assigned and not wrongly assigned as the same

cluster as a homozygous cluster. These three quality control parameters (GC score, GT score and Cluster Sep) are measurements that range from 0 to 1 with 0.15 as a lower limit for quality genotyping.

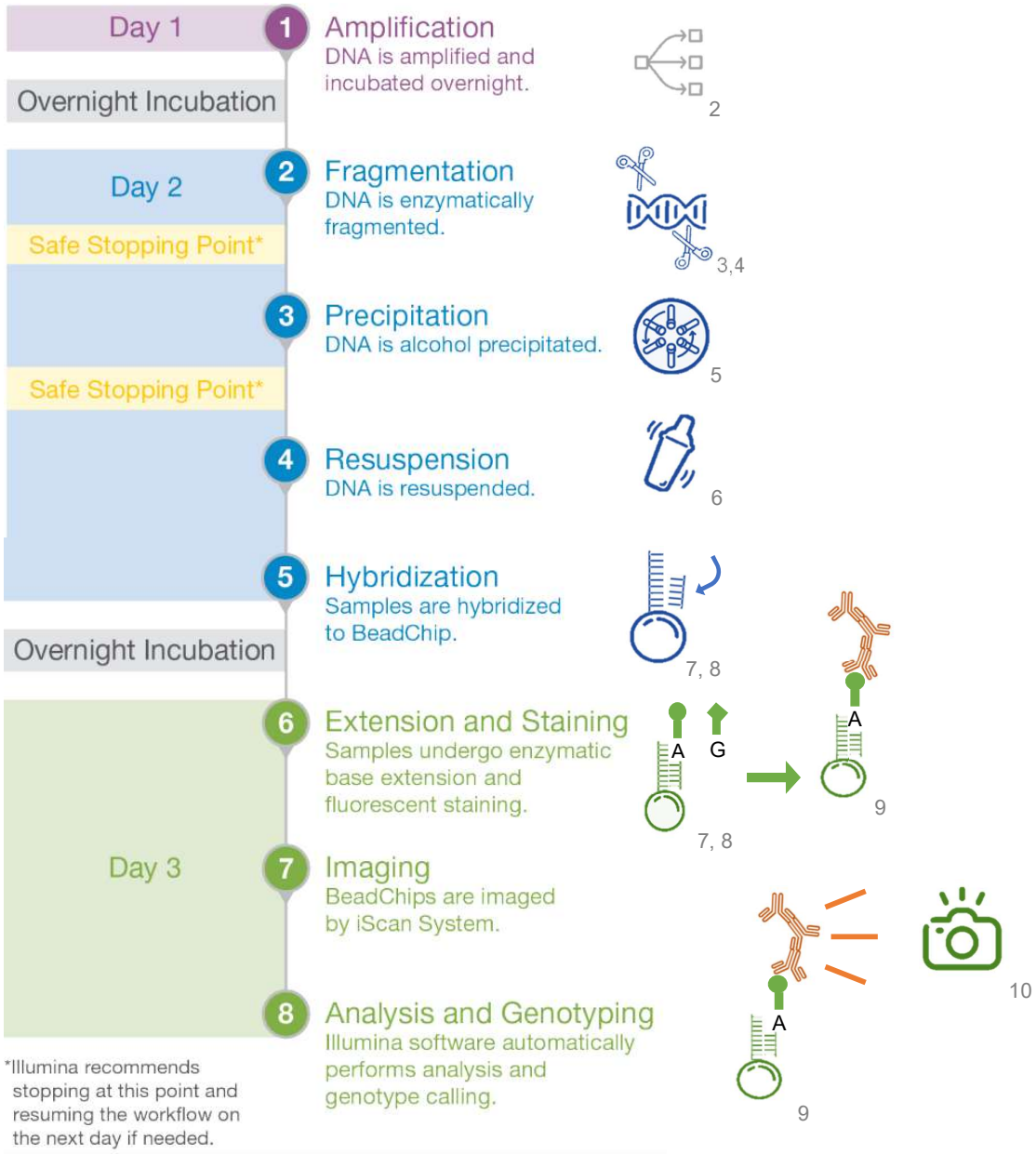


Figure 3.10 Infinium Beadchip Workflow

Adapted from Illumina (2018) and the Noun project

After quality control of the data the allelic calls were exported. Different notations are used by Genome Studio for this. The PLUS strand notation was used since it corresponded with the GSA v 2.0 chip SNP lists and was also compatible with Ensembl and NCBI and Fast references.

Among the multiple genes and SNPs involved with oestrogen biotransformation, the genes encoding the enzymes involved in the first step of oestrogen biotransformation (i.e. *CYP1B1, CYP3A4, CYP1A1* and *CYP1A2*) as well as the enzyme *COMT* which catalyses the subsequent reaction (methylation of the hydroxy oestrogens), were selected for analyses in this study. This selection of genes resulted in a collection of 250 SNPs which had to be filtered for further analyses.

3.6 SNP selection and Haplotype Linkage

To achieve both aims of the study, group specific association analyses had to be done. Therefore, the samples were grouped into African, Caucasian and Combined Controls study groups for further SNP selection. The SNP selection was done per group and also per chromosome (since some genes are in close proximity).

The extracted genotyping data (Genotype Calls) from Genome Studio 2.0 was rearranged into .ped files and .map files for the importation into PLINK. The data in the .ped files was arranged under the following headings: FID, and IID (Sample names), PID, and MID (paternal and maternal knowledge as zero), Gender (woman was assigned the number two), Phenotype (within this context was zero) and then numerically the SNP depending on the location on the chromosome per chromosome. Within the .map file the data was arranged under the headings Gene, rsID (SNP name), and the location of the SNP on the chromosome. The files were re-coded to .ped and .info files by PLINK using the <-- recode HV> command line. These files were then analysed in the Haploview program.

Haploview software was used to calculate a number of parameters that determined whether a specific SNP could be included for further association analyses. These parameters included the minor allele frequency (MAF), the Hardy-Weinberg equilibrium (HWE) population allele frequency *p*-value, the difference between the observed heterozygosity (ObsHET) and the predicted heterozygosity (PredHET) score (calculated using the chi-square test (x^2)), and the number of minor allele carriers (heterozygous or homozygous for the minor allele-must at least carry one minor allele) within a group (Barret *et al.,* 2005). Finally, for the purpose of association analyses with metabolite levels, every SNP within a population group should also

contain sufficient variation in terms of allele frequencies to be able to associate with metabolite of interest.

Within Haploview all the SNPs that demonstrated no Observed Heterozygosity were removed. All SNPs that had a significant ($\alpha \leq 0.05$) x^2 between the Observed and Expected Heterozygosity, as well as significant HWE *p*-values were removed since the significant result might indicate that another factor might be influencing the differences in SNP frequencies. Any associations, determined within SNPs deviating from HWE, could be the result of numerous HWE instead of genotype causing a phenotype change. Lastly, according to de Winter (2009) and Hertzog (2008) a minimum of 10 samples for an associating analysis must be obtained for credible results. Therefore, SNPs for which there were not a minimum of 10 minor allele carriers (either heterozygosity or homozygosity for the minor allele) within a group were excluded (i.e. where the combined number of samples that carried one or two minor alleles was not at least 10 samples, the SNP was excluded due to a lack of power for association analysis). By applying all these criteria, the SNPs per study group were selected for further analyses.

Haploview calculated haploblocks by determining if the frequencies of the SNPs that are located close to each are random or non-random. When the frequencies are above 70 % as non – random linkages, then those variants will demonstrate as a haploblock. This means that these SNPs have a high probability of being inherited together. These haploblocks were also linked to each other through the same frequency calculation yielding a factor known as Hendricks multiallelic D'. The degree of linkage is indicated as a percentage and corresponds with the thickness of the line connecting the haploblocks demonstrated on the schematic example generated from this study (Barrett *et al*., 2005) (Figure 3.11).

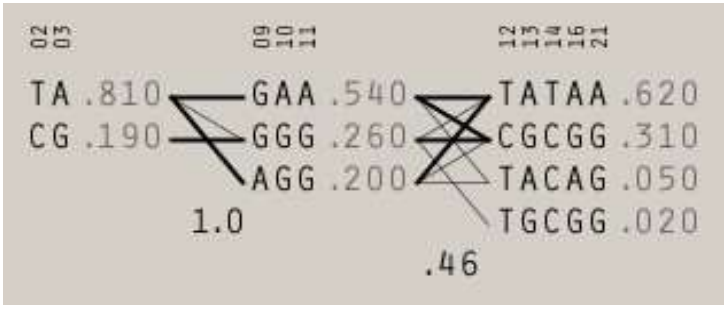


Figure 3.11 Haplotype display of Multiallelic D' line-thickness

3.7 Phenotypic Data Quality Control

From this point onward, the study groups were divided into 4 groups namely, African Controls (23 samples), Caucasian Controls (25 samples), Combined Controls (48 samples), and the COC users (25 samples). The SNPs selected for the Caucasian group (in 3.5 above) were included in both the Caucasian Controls and the COC users group, given that they share the same ethnicity and should theoretically have the same genotype frequencies.

The urinary concentrations of 33 hormones or oestrogen metabolites and the levels of three blood oxidative stress markers (ROS, FRAP GSHt) were included in the phenotypic data set (Table 3.2). This dataset was generated during the eBOSS study from either high-performance liquid chromatography (HPLC), liquid chromatography mass-spectrometry (LC-MS) readings and the ROS is represented by serum peroxide measurements.

All the metabolite data was $\log_{10}$ transformed after which outliers were removed if they were more than three times the standard deviation removed from the mean (about none to two samples were removed). This order of calculations was followed in order to retain as many samples as possible from this small sample size study. The metabolites were tested for normal distribution within the program IBM SPSS using the Shapiro Wilk test which is considered the most accurate when the sample sizes are below 50. Although non-parametric tests are usually suggested for non-normally distributed data, (as was the case with this data set even after $\log_{10}$ transformation), general linear models tolerate moderate violations to normal distribution (Garson, *et al.,*2012). A skewness of ± 2 and a kurtosis of ± 7 was, therefore, considered acceptable for this study (Byrne & Vijver, 2010).

Table 3.2 List of oestrogen, oxidative stress, and biotransformation metabolites/markers

Metabolites			
Oestradiol-3-glucuronide	16-Ketoestradiol	4-Methoxyoestrone	D-2-hydroxyoestrone-6N3-adenine
Oestradiol-17-sulphate	17-Epioestriol	Androstenedione	D-2-hydroxyoestradiol-6N3-adenine
Oestradiol-3-sulphate	2-Hydroxyoestradiol	Oestradiol Alpha	D-4-hydroxyostrone-1N7-guanine
Oestriol-3-sulphate	2-Hydroxyoestrone	Oestradiol Beta	D-4-hydroxyoestrone-1N3-adenine
Oestriol-16-glucuronide	2-Methoxyoestradiol	Oestriol	D-4-hydroxyoestradiol-1N7-guanine
Oestrone-3-sulphate	2-Methoxyoestrone	Oestrone	D-4-hydroxyoestradiol-1N3-adenine

Oestrone-3-glucuronide	4-Hydroxyoestradiol	Oestrone-2-hydroxy-3-methyl ether	Reactive oxygen species (ROS)
16-Epiestriol	4-Hydroxyestrone	Progesterone	Ferric reducing ability of plasma (FRAP)
16-Hydroxystrone	4-Methoxyestradiol	Testosterone	Total red blood cell glutathione (GSHt)
Legend			
N - Nitrogen			

3.8 Association Analysis

After SNP selection and processing of the metabolite data, preliminary association analysis was done in PLINK. Significant SNP-metabolite associations were then further analysed using SPSS software. The results were presented in tables as well as multiple Boxplots (metabolite vs all the relevant SNPs) to illustrate the nature of association and compare it between study groups.

3.8.1 PLINK

The PLINK software was fast to use and highlighted the SNP-metabolite associations to be studied in the IBM SPSS statistics software. The .ped and .map files were used for analysis in PLINK along with a metabolite text file (or phenotype file) also containing the headings FID, and IID (Sample names) and then a column for each metabolite. The following PLINK command lines where used: <--logistics>

- <--allow-extra-chr>
- <--pheno>
- <--all-pheno>
- <--adjust>
- <--freq

The logistics command is more flexible with non-normally distributed data or small sample sizes, although this specific command might take a very long time to process (Anderson *et al*, 2010). Significant SNP-metabolite associations were determined by the False Discovery Rate of Benjamini Hochberg (FDR-BH) adjusted p-value. A significant p-value of below 0.05 was highlighted as a strong association and a *p*-value of 0.05-0.08 was considered a possible association. The latter were also taken note of since small sample sizes may not have the power to detect less prominent associations and association with a *p*-value in this range might contain false negative results. Since correction for multiple testing FDR-BH is known to have

significant *p*-values for single sample (rare allele) associations. (Zhang *et al.,* 2002), the results had to be further studied in SPSS. PLINK also shows only that an association exists and does not elucidate which genotype specifically associated with the metabolites (Zhang *et al.,* 2002).

3.8.2 SPSS

The SNP-metabolite associations highlighted by PLINK preliminary association analyses are further analysed in great detail in SPSS. The SPSS software was used to determine which allele groups were significantly associated with metabolite levels and also to determine the power of the association.

A single excel sheet that contained all the SNPs and metabolites that were associated in PLINK as well as the covariates was created to import in SPSS. The covariates that could also influence oestrogen metabolites and oxidative stress related metabolites were age, BMI, the menstrual phase and where applicable, ethnicity.

When analysing a SNP, it must be noted that there is a major allele (M) and a minor allele (m) and that human individuals are diploid in nature and they carry two alleles, therefore many allele models have been created to study SNPs (Kirpich *et al*., 2018). The major allele was determined by the Ensembl (1000 Genomes) database "All" category where all populations accumulated data was used to determine the major and minor allele. In order to determine which genotype drives the association, three allele models focusing on the variant / minor allele-perspective can be applied. The first model is the Additive (co – dominant) model in which the three different genotypes are treated as separate groups to be compared (i.e. homozygous wild (MM), heterozygous (Mm) and homozygous variant (mm)). The second model is Dominant model (MM and Mm / mm) in which the heterozygous and homozygous variant is grouped together for comparison with the homozygous wild genotype to determine if the variant has a dominant effect. The last model is the recessive model (MM / Mm and mm) in which the homozygous wild type and heterozygous is grouped together for comparison to the homozygous variant. This was done for each SNP-metabolite association.

The *p*-values of the Games-Howell post hoc test were used to establish which allele groups (MM, Mm, mm) within the Additive Model for a specific SNP significantly differed from one another. The G-H post hoc is one of the more powerful *post hoc* tests that is also tolerant towards uneven and small sample sizes (Fields, 2009).

A non-parametric test namely Man-Whitney (MW), was also performed and the results were compared to that of the general linear model (GLM). The GLM results did not significantly differ from the MW results and, thus the most convenient and informative test was used for association analysis of this study. The GLM also calculated the F statistics that can be used

for interpretation alongside the p – value. Lastly, the GLM also calculated the partial squared eta (η^2) that was used to determine the power of this specific association. All these analyses were done with all the three Allele Models.

3.8.3 Discussion and Graphical Presentation of Results

The SNPs involved in significant association results were selected to determine, through literature, the effect this polymorphism had the respective enzyme activity and expression. The literature articles of interest were chosen by searching within google scholar with the three quoted (“”) keywords comprising of the SNP-ID, metabolite and study group and linked with the capitalized “AND” within the search bar. When this search yielded no articles that contained all three these keywords, separate searches containing either SNP-ID and metabolite or SNP-ID and study group were done where the information of those articles were combined to conclude a possible effect or function.

All the significant associations were graphically presented in boxplots of metabolite concentration versus all associated SNPs for all group in order to make comparisons between individuals as well as between different ethnic groups. Some of the metabolites or SNPs have been highlighted to indicate a need for future study as potential biomarkers.

CHAPTER 4 – RESULTS

Correlation analyses between genetic data and metabolite data might shed light on the effect which specific genetic variations (genotype) could have on the metabolic activity (phenotype). For genotype analysis, firstly, the quantity and integrity of the DNA samples were checked. Secondly, SNPs were selected based on the lower Minor Allele Frequency (MAF) limits, calculated in Haploview (4.1) and preliminary association analyses in PLINK (1.9b6.22 32 bit). Thirdly, the metabolic data was transformed to match the limitations and assumptions of mathematical selection and association tests. Through association analyses in PLINK and SPSS (28.0.0.1) specific SNPs could be linked with specific metabolites. Group comparisons of all significant associations were made to reveal the effect of genotype on metabolite levels. The results obtained indicate that certain genes or SNPs may have been understudied in African populations, and also highlight potential biomarkers or SNPs that may be added to risk score determination.

4.1 gDNA Isolation and Quality Control

The quality and purity of the isolated genomic DNA (gDNA) was examined through spectrophotometric and electrophoresis methods. For **spectrophotometric** absorbance checks, the A260/A280 ratios acquired from the Nanodrop One / OneC Microvolume UV – Vis Spectrophotometer (81 Wyman Street, Waltham, Massachusetts 02451 USA, ThermoFisher Scientific) for all of the 74 samples, were between 1.7 and 2.1, indicating acceptable purity (Table 4.1). Most A260/A230 ratios were within the range of 1.5 to 3.0 as requested by Centre for Proteomic Genomic Research (CPGR) based in Cape Town; with eight of the 74 samples below the acceptable range, indicating possible guanidine (chaotropic salt) contamination (Boesenberg-Smith *et al*., 2012) from the extraction steps. Guanidine contamination could interfere with genotyping methods which rely on DNA–protein interaction (Boesenberg-Smith *et al*., 2012; Madhad & Sentheil, 2014). However, the Infinium Global Screening Array v2.0 (GSA v 2.0) Beadchip and Illumina iScan system do not rely on DNA-protein interactions and therefore, were minimally affected by guanidine contaminations (Adler *et al*., 2013). Consequently 74 gDNA isolated samples were deemed pure enough for genotyping.

CPGR (2022) requested 500 ng of gDNA at a concentration of 50 ng/µℓ for each sample. All samples measured by the Invitrogen Qubit 2.0 Fluorometer instrument (5791 Van Allen Way Carlsbad, California, 92008 Life Technologies), were provided as such. For sample eBOSS032, only 450 ng of DNA at 50 ng/µℓ could be provided. This was considered a

sufficient quantity, since Adler *et al.* (2013) have shown that typically 200 ng DNA (or slightly more for low quality samples) were enough for successful genotyping using the Infinium GSA v 2.0 Beadchip. All of the samples were of sufficient purity and quantity for GSA v2.0 Genotyping, as shown in the appendix Table S0.1

Table 4.1 74 Samples Spectrophotometry, Fluorometer Ranges and Concentrations

Range	Average A260/A280	Average A260/A230	Concentration (ng/µℓ)	Supplied Weight gDNA (ng)
Min	1.79	1.29	49.90	456.00
Max	2.00	2.21	50.61	790.40

Legend

The table summarizes the gDNA integrity quality control results and amount of DNA that was sent to CPGR as minimum and maximum values of all 74 samples. The absorbance ratios gave an indication of the purity of the gDNA. The minimum and maximum concentrations measured for the samples are also listed. The last column indicates minimum and maximum amount gDNA/sample that were sent for genotyping. All samples were sent for genotyping, although some samples were low in quantity or had possible salt contamination that may influence genotyping efficiency.

CPGR (2022) requested visual confirmation of the presence of high molecular weight gDNA in the samples. Thus, in order to detect small fragment contamination, a 2% agarose gel **electrophoresis** of the gDNA samples was done of which the results are shown in Figure 4.1 (b-c). Since originally 35 of the 74 samples [non-filled squares in panel (a)] showed guanidine salt contamination, gDNA isolation from these samples were repeated, to improve the overall purity. Gel electrophoresis was repeated for these samples as well [See Fig. 4.1 panel (c)], gel 4). Representative images from two gels are shown in Fig. 4.1 (b) and (c). Images of all the complete gels are given in appendix Figure S0.1.

The samples loaded in horizontal Row 2 (Figure 4.1 b) showed intact, bright bands at the top of the gel indicating respectively whole, concentrated and high molecular weight DNA. The high molecular weight (17kb) ladder confirmed the presence of gDNA as the bands were closer to the wells than the position corresponding to the upper 17 Kb band of the ladder. The sample bands are shown as a single fragment indicating intact, non-degraded DNA. The yield of DNA is indicated by the high intensity of the band and was considered to be sufficient.

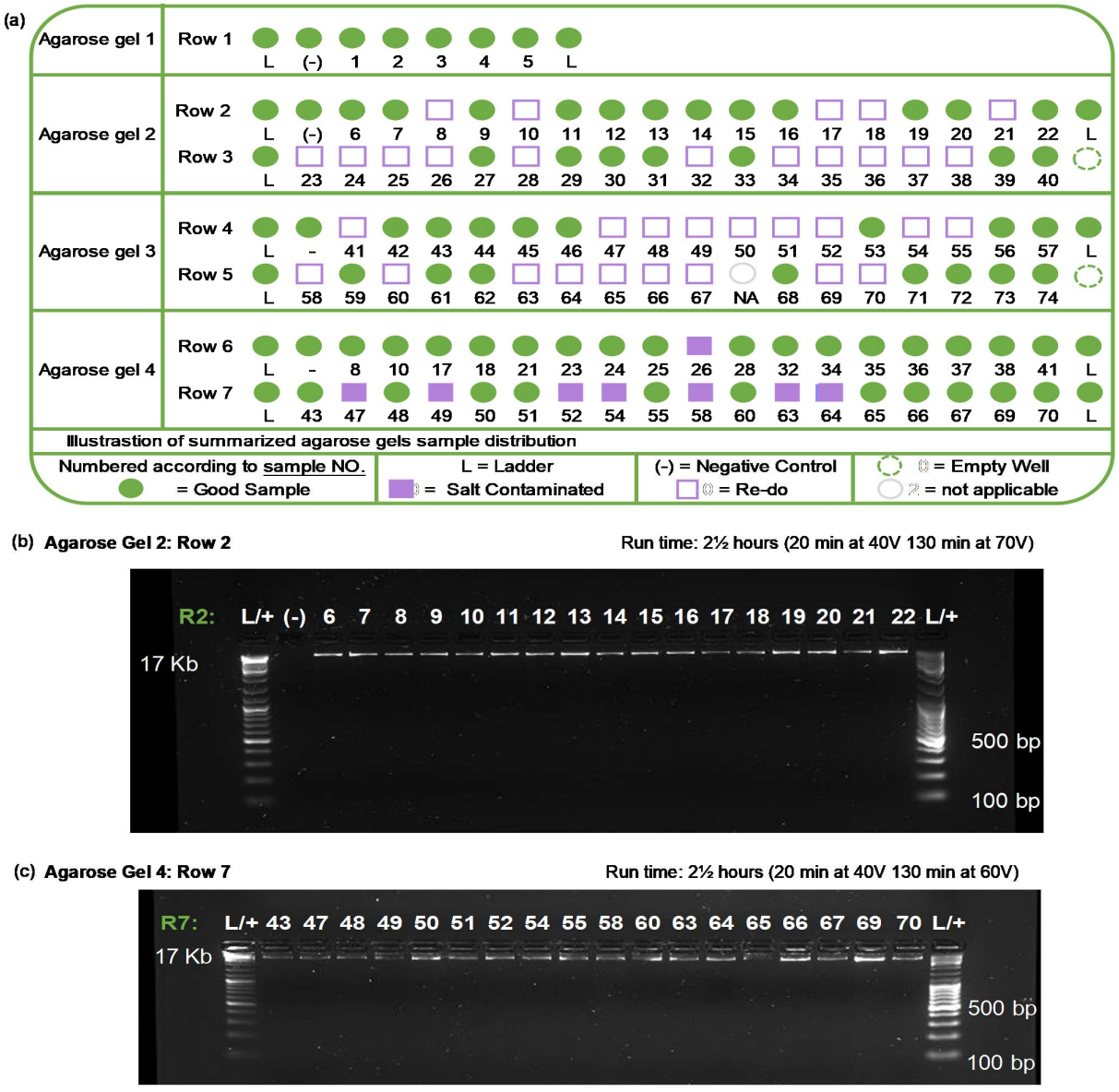


Figure 4.1 2% Agarose gel electrophoresis of gDNA with a sample distribution illustration

Legend

(a) Illustration: Gel 1-3, all samples indicated by a filled circle or square were of acceptable quality to be used in this study. All samples indicated as open squares were selected and re-isolated to improve quality. These samples were analysed again and are shown in gel 4 as filled shapes.

(b - c) Parts of gels 2 and 4 are shown. The first column in every row was a No-limits 17 Kb ladder. The last column contained a 100 bp ladder (SM0241). Row 2 (b) second column was the negative control. On Gel 2, Row 2 samples NO. 6-22 (eBOSS008-028) were loaded, and on Gel 4, Row 7 some of the re-isolated samples [NO.43 (eBOSS050) to sample NO.70 (eBOSS093)] were loaded in numerical order.

The absence of bands at the height of the 100 bp ladder fragment indicates that the samples were free of RNA of at least 100 bp or longer in length. Samples in horizontal Row 7 (Figure 4.1 c) revealed lower DNA yields but still demonstrated intact gDNA that remained enough for genotyping purposes. The spectrophotometry, fluorometry and electrophoresis results meet the requirements of objective 3.

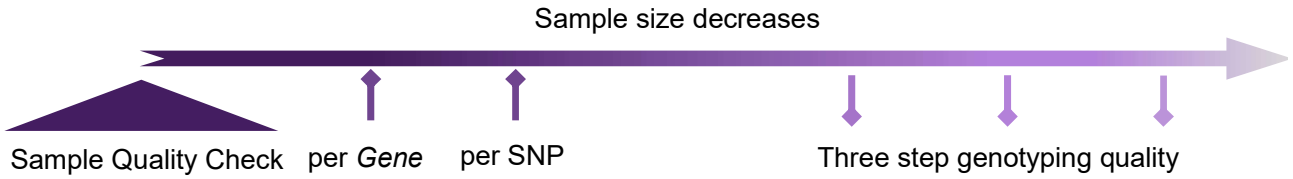
4.2 Genotyping Quality Control

The genotyping results (objective 4) for all of the 74 samples were obtained from CPGR and subjected to quality control (QC) using the program GenomeStudio 2.0. Although the QC was done for all SNPs included in the array, only the SNP results for the selected genes of interest in this study, are presented and discussed here.

The detailed QC results for all samples can be found in the appendix Table S0.2, with a summary of this data presented here in Table 4.2 as obtained from GenomeStudio 2.0 software. As seen in Table 4.2 and Table S0.2, the allele frequency calls for 73 of 74 samples was 98-99%, indicating high genotyping efficiency and no degradation or contamination (Flickinger *et al.,* 2015; Wiggans *et al*., 2010). One sample (eBOSS052, Sample NO. 44) from the African Control group had a low call rate of 82.6%. According to Wiggans *et al*. (2010) a low call frequency usually indicates poor quality or low quantity of DNA and was therefore excluded from further analyses. In addition to the allele frequency call, for each SNP, the GenCall (GC) Score, GenTrain (GT) Score and Cluster Separation scores were determined. The scores function on a scale from 0 to 1, where 0 indicates unsuccessful genotyping, clustering or cluster separation and where 1 indicates success in those fields.

In Table 4.2, three SNPs of gene *CYP1B1* (rs1800440, rs10012 and rs2551188) were selected and listed to demonstrate genotyping quality controls that all 179 SNPs were subjected to. For the rs10012 SNP specifically, another two samples (eBOSS012, Caucasian Control and eBOSS070, African Control) were removed from the study due to low GC scores (< 0.15), which would influence accurate genotyping. The GC scores for 73 samples for the remaining SNPs of interest were high (above 0.15; Table 4.2 and S0.2), indicating that all the samples were close enough to a cluster to be assigned a genotype. All the SNPs with a GT Score of 0.80 confirmed the clustering of the genotypes were successful, since the closer the value is to 1.00, the higher the clustering quality is (Zhao *et al.,* 2018). SNPs rs1800440 and rs10012 had a GT score below 0.80. After manual reviewing as shown in Figure 4.2, it was determined that these three SNPs had successfully clustered into three distinct genotypes, especially for a small sample size study (Hertzog, 2008).

Table 4.2 74 Samples Summarized Allele Calls QC within GenomeStudio 2.0



Call Frequency		Gene	SNPs	Sample	GC Score	GT Score	Cluster Sep
min.	0.99	*CYP1B1*	rs1800440	All	0.81	0.79	1.00
max.	0.98		rs10012	eBOSS0012	0.00	0.44	0.37
eBOSS052	0.83			eBOSS0070	0.00		
				min.	0.16		
				max.	0.23		
			rs2551188	All	0.80	0.80	1.00

Legend

Call Freq - Call Frequency; CYP - Cytochrome P450; COMT - Catechol - O - methyltransferase;

GC Score - GenCall score; GT Score - GenTrain score; Cluster Sep - Cluster Separation score;

All - All the samples; The Rest - Remainder of Samples

The illustration above the table demonstrated the order of genotyping results and quality control. Per sample, all the 74 samples collective successful genotyping percentages are shown in the Call Freq columns. All purple highlighted samples within this table were removed from further association analysis. The genes were arranged according to chromosome number. The SNPs were arranged according to the position on the chromosome. Only SNPs that significantly associated are listed within this table. The Sample Names were arranged according to an increasing GC score. The GC score is a parameter that determines whether a genotype can be accurately assigned to the specific sample. The GT score and Cluster Sep are parameters that determine proper genotype cluster formation and if the cluster are distinct from one another respectively.

Most SNPs' Cluster Separation (Cluster Sep) score was as high as it can be, indicating that the clusters were definitively separated well enough to ensure separate genotypes (Zhao *et al*., 2018). Although SNP rs10012 had a low non–optimal Cluster Sep score of 0.33, the cluster formation and separation was deemed sufficient after manual reviewing as seen in Fig 4.2 c. that represents the removed sample 44 (eBOSS052).

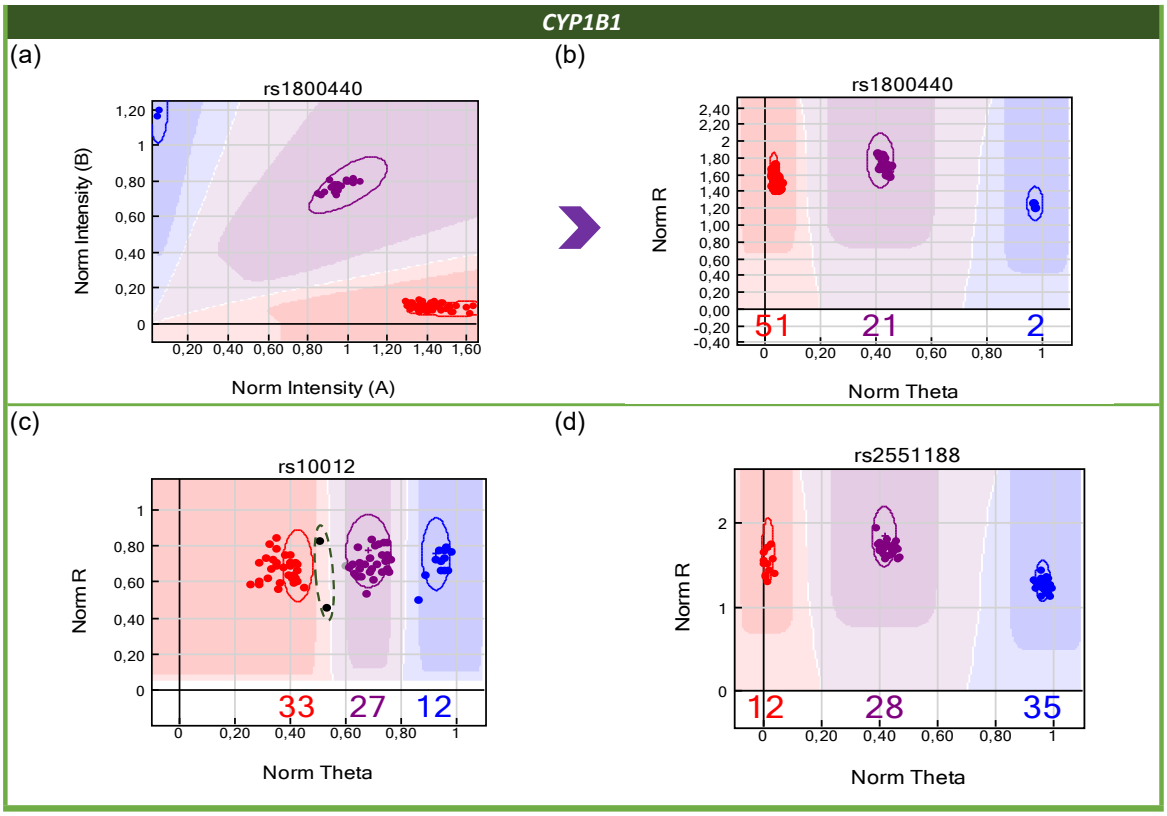


Figure 4.2 Illustration of CYP1B1 Allele Call Normalization within GenomeStudio 2.0

Legend

Norm Intensity A/B *- Measurement of fluorescent light assigned to alleles labelled as A/B*

Norm R *- Norm Intensity A + Norm Intensity B*

Norm theta (Θ) *- Formula based on heterozygous representation (0.2 < heterozygous > 0.8)*

CYP - Cytochrome P450

The graphs display the normalised fluorescence intensity recorded for each allele of selected SNPs. As an example, the normalised intensities for alleles A and B of rs1800440 are plotted against each other in (a) and the combined intensities against the normalised theta value in (b) and the latter is repeated for the other SNPs (c-d). The relevant SNP IDs are given at the top of each graph. For all the graphs displayed the red colouring represents homozygous AA alleles, purple represents heterozygous AB alleles, and blue represents homozygous BB alleles. The homozygous AA alleles (red) were located to the left of the graph and the homozygous BB alleles (blue) to the right with heterozygous AB alleles (purple) in the middle for (b-d). Alleles A and B are determined by the GSA chip v2.0 setup. Within all these graphs, although not always visible, is a grey dot that represents sample eBOSS052 that was excluded.

Figure 4.2 graphically displays the allelic calls and the normalization from the numerical parameters listed in Table 4.2. Here it is important to determine if the 73 sample dots are 1)

located within or near a cluster, 2) if a cluster is properly formed, and 3) if the clusters are separated well enough. Fig 4.2 (a-d) contains a single grey dot (which is not always visible) Fig. 4.2 contains the three *CYP1B1* SNPs (rs1800440, rs10012 and rs2551188) and SNP rs1800440 is an excellent example of complete successful genotyping (Figure 4.2 a and b).

The SNP rs10012 [Fig 4.2 (b)] was almost fully genotyped - only the degraded sample and the two non-genotyped samples (eBOSS012 and eBOSS070) highlighted by a dashed green oval is shown not to belong to a genotype cluster (c). SNP 2551188 (d) was similar to the example rs1800440 (Fig 4.2 a-b). All 179 SNPs were manually reviewed according to Illumina (2010) guidelines as demonstrated in Fig 4.2 concluding objective 5.

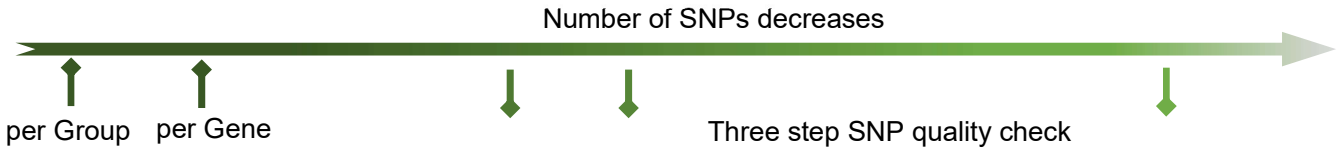
4.3 SNP Selection Parameters

The next step in preparation for association analysis entailed a QC step to identify SNPs in Haploview that met the minimum requirements for further association analysis. Although association analyses were done for the four groups that are shown in Tables 3.1, the two groups (Caucasian Control and COC users) were combined into a single Caucasian group during this SNP selection step (Table 4.3). Caucasian Controls and COC users have the same genetic background and are consequently, combined into a collective group named 'Caucasian'. This combination is to improve the accuracy of the quality control results adapted from the minor allele frequency as much as possible with these small sample sizes. The QC results for Caucasian were used for both Caucasian Controls and COC users during the SNP selection step. The Combined Controls merged all the African Control and Caucasian Control samples.

The criteria for SNP selection were based on the Hardy–Weinberg equilibrium (HWE) population allele frequency *p*-value of <0.05, a non-significant difference in the observed heterozygosity (ObsHET) and predicted heterozygosity (PredHET) determined by a Pearson chi square test (x^2), and a minimum of 10 minor allele carriers (either heterozygous or homozygous for the minor allele) within a group. The purple highlighted SNPs in the "x^2", "HWE-*p*val" and "minor allele carriers" columns were the SNPs that had to be removed from that population group for association analysis. Minor allele carriers are samples that were either heterozygous or homozygous for the minor allele. The significant results from one of the selection parameters (x^2) can be found in the appendix in Table S0.3.

The Tables 4.3-4.5 depict the possible significant SNPs that were selected per population group respectively, 1) Caucasian group, 2) African group, and the 3) Combined Controls group. The Caucasian group consisted of Caucasian Controls and COC users for which the respective number of samples and minor allele carriers are indicated in this order under the "Sample Size" and "Minor Allele Carrier" columns of Table 4.3.

Table 4.3 Caucasian population group SNP selection and Quality Control within Haploview



Population	Gene	SNPs	x^2	HWE - *pval*	▼MAF▼	Sample Size (control/COC)	Minor allele carriers (control/COC)	Alleles
Caucasian	*CYP1B1*	rs1056837	0.08	0.01	0.33			
		rs1056836	0.08	0.01	0.33			
		rs10012	0.48	0.36	0.33	24 / 25	12 / 13	G:C
		rs2551188	0.47	0.33	0.32	25 / 25	12 / 13	C:T
		rs1800440	0.66	0.71	0.26	25 / 25	11 / 12	T:C
		rs10916	0.69	0.60	0.20		9 / 8	
		rs162562	0.68	0.58	0.20		9 / 8	
		rs55989760	1.00	1.00	0.01		1 / 0	
		rs57865060	1.00	1.00	0.01		0 / 1	
		The Rest	0.00					
	CYP3A4	rs2246709	0.12	0.01				
		rs4646437	0.81	1.00	0.10		6 / 3	
		rs2242480	0.83	1.00	0.09		4 / 5	
		rs3735451	0.84	1.00	0.09		4 / 5	
		rs6956344	0.84	1.00	0.09		4 / 5	
		rs2687116	0.93	1.00	0.05		4 / 1	
		rs35599367	0.93	1.00	0.05		3 / 2	
		rs28988600	0.98	1.00	0.02		2 / 0	
		rs34642455	0.98	1.00	0.02		1 / 1	
		rs28988574	1.00	1.00	0.01		1 / 0	
		The Rest	0.00					
	CYP1A1 / CYP1A2	rs2472304	0.39	0.33	0.38	25 / **25**	15 / **18**	
		rs2470890	0.39	0.33	0.38	25 / **25**	15 / **18**	
		rs762551	0.76	0.91	0.31	25 / 25	14 / 13	A:C
		rs2606345	0.56	0.55	0.30	25 / 25	12 / 15	A:C
		rs4646421	0.89	1.00	0.07		3 / 4	
		rs3743484	0.96	1.00	0.04		2 / 2	
		rs1048943.3	0.97	1.00	0.03		1 / 2	
		rs1799814	0.97	1.00	0.03		3 / 0	
		rs4646427	0.98	1.00	0.02		0 / 2	
		rs2069526	0.98	1.00	0.02		0 / 2	
		rs41279188	1.00	1.00	0.01		1 / 0	
		rs17861094	1.00	1.00	0.01		0 / 1	
		rs4646422	1.00	1.00	0.01		0 / 1	
		rs4646425	1.00	1.00	0.01		0 / 1	
		rs56276455	1.00	1.00	0.01		1 / 0	
		The Rest	0.00					
	COMT	rs5746849	0.35	0.24	0.46	**25 / 25**	**17 / 16**	
		rs740603	0.30	0.18	0.46	**24 / 25**	**16 / 16**	

		rs5993883	0.47	0.39	0.45	25 / **25**	15 / **18**	
		rs4633	0.93	1.00	0.38	25 / **25**	13 / **18**	
		rs4680	0.93	1.00	0.38	25 / **25**	13 / **18**	
		rs6269	0.83	0.90	0.33	25 / **25**	11 / **16**	
		rs2239393	0.83	0.90	0.33	25 / **25**	11 / **16**	
		rs4646312	0.76	0.91	0.31	25 / **25**	11 / **16**	
		rs4818	0.76	0.91	0.31	25 / **25**	11 / **16**	
		rs933271	0.42	0.24	0.28	25 / 25	11 / 11	T:C
		rs165599	0.52	0.46	0.27	25 / 25	13 / 12	A:G
		rs174696	0.96	1.00	0.25	25 / 25	10 / 12	T:C
		rs3819619	0.41	0.17	0.20		9 / 7	
		rs737866	0.93	1.00	0.19		8 / 9	
		rs737865	0.93	1.00	0.19		8 / 9	
		rs35788262	0.87	0.98	0.18		9 / 7	
		rs4646316	0.87	0.98	0.18	**25** / 25	**6** / 10	C : T
		rs9332377	0.66	0.70	0.14		5 / 9	
		rs174699	0.97	1.00	0.03		2 / 1	
		rs45616631	0.98	1.00	0.02		0 / 2	
		rs165631	0.98	1.00	0.02		2 / 0	
		rs34068292	1.00	1.00	0.01		0 / 1	
		The Rest	0.00					

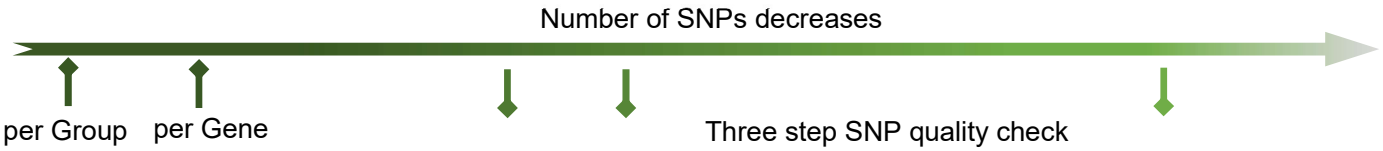
Legend

Caucasian - Caucasian Controls and COC users	HWE - Hardy Weinberg equilibrium
COC – Combined oral contraceptives users (Caucasian)	MAF - Minor allele frequency
CYP - Cytochrome P450	*pval - p-value*
COMT – Catechol-O-methyltransferase	*Minor allele carrier - Minimum of 10 samples either*
The Rest – The remaining SNPs	*heterozygous or homozygous for the minor allele*
x^2 - Chi Square test	*Coloured/Highlighted SNPs were removed*

This is the SNP Selection and QC for the Caucasian population group which is the combination of Caucasian Controls and Combined oral contraceptive (COC) users. The genes were arranged according to the chromosome number. The SNPs were arranged according to decreasing MAF. All the SNPs that had no observable minor alleles were removed. The x^2 (calculated by the ObsHET and PredHET) and HWE served as mutation interference QC parameters (cut-off $p < 0.05$). The Minor allele carriers determines the minimum of 10 samples needed for sufficient association analysis not only for the combined homozygous minor genotype and heterozygous genotypes but also the homozygous wild genotype.

The purple highlighted and bolded values are the SNPs that were not selected according to the three criteria listed on page 59 for one specific Caucasian group. With that in mind, it concludes that the Caucasian Control group had 17 SNPs that were selected, and the COC group had 9 SNPs that were selected for association analysis.

Table 4.4 African population group SNP selection and Quality Control within Haploview



Population	Gene	SNPs	x^2	HWE - *pval*	▼MAF▼	Sample Size	Minor allele carriers	Alleles
	CYP1B1	rs10916	0.08	0.14	0.41	23	13	A:C
African		rs10012	0.25	0.40	0.41	22	13	G:C
		rs2551188	0.22	0.33	0.39	23	13	C:T
		rs162562	0.47	0.62	0.30	23	11	T:G
		rs1056837	0.78	1.00	0.11		5	
		rs1056836	0.84	1.00	0.09		4	
		rs4986889	0.91	1.00	0.07		3	
		rs4987134	0.94	1.00	0.04		2	
		rs4986888	1.00	1.00	0.02		1	
		The Rest	0.00					
	CYP3A4	rs6956344	0.95	1.00	0.33	23	13	C:T
		rs2687116	0.65	1.00	0.26	23	11	A:G
		rs2246709	0.66	0.28	0.26	23	12	C:A
		rs4646437	0.02					
		rs3735451	0.42	1.00	0.17		7	
		rs2242480	0.18	1.00	0.17		7	
		rs4646440	0.45	1.00	0.09		4	
		rs28371757	0.49	1.00	0.02		1	
		rs4986907	1.00	1.00	0.02		1	
		The Rest	0.00					
	CYP1A1 / CYP1A2	rs4646421	0.91	1.00	0.37	23	14	
		rs762551	0.71	1.00	0.33	23	13	G:A
		rs17861157	0.52	0.96	0.17		8	
		rs2069526	0.65	0.83	0.15		6	
		rs17861094	0.54	0.62	0.13		5	
		rs4646427	0.54	0.62	0.13		5	
		rs4986883	0.44	0.43	0.11		4	
		rs2472304	0.78	1.00	0.11		5	
		rs17861084	0.84	1.00	0.09		4	
		rs2606345	0.94	1.00	0.04		2	
		rs4986884	1.00	1.00	0.02		1	
		rs4986881.1	1.00	1.00	0.02		1	
		rs17861152	1.00	1.00	0.02		1	
		rs45540640	1.00	1.00	0.02		1	
		rs2470890	1.00	1.00	0.02		1	
		The Rest	0.00					
	COMT	rs3819619	0.51	0.89	0.46	23	17	
		rs2239393	0.26	0.54	0.44	23	17	

		rs933271	0.08					
		rs5993883	0.04					
		rs6269	0.66	1.00	0.39	23	15	
		rs740603	0.12	0.18	0.37	23	12	G:A
		rs174696	0.51	0.89	0.35	23	14	
		rs9332377	0.19	0.26	0.33	23	11	C:T
		rs4646316	0.71	1.00	0.33	23	13	C:T
		rs5746849	0.89	1.00	0.28	23	11	G:A
		rs4633	0.89	1.00	0.28	23	11	C:T
		rs4818	0.71	0.93	0.26	23	10	C:G
		rs737866	0.55	0.71	0.24		9	
		rs737865	0.55	0.71	0.24		9	
		rs165599	0.09	0.07	0.22		7	
		rs4680	0.34	0.57	0.22	23	10	G:A
		rs4646312	0.78	1.00	0.17		7	
		rs769224	0.54	0.62	0.13		5	
		rs8192488	0.31	0.32	0.11		3	
		rs35788262	1.00	1.00	0.02		1	
		rs34068292	1.00	1.00	0.02		1	
		rs174699	1.00	1.00	0.02		1	
		The Rest	0.00					

Legend

African - African Controls

CYP - Cytochrome P45

COMT – Catechol-O-methyltransferase

The Rest - The remaining SNPs

x^2 - Chi Square test

HWE - Hardy Weinberg equilibrium

MAF - Minor allele frequency

pval - p-value

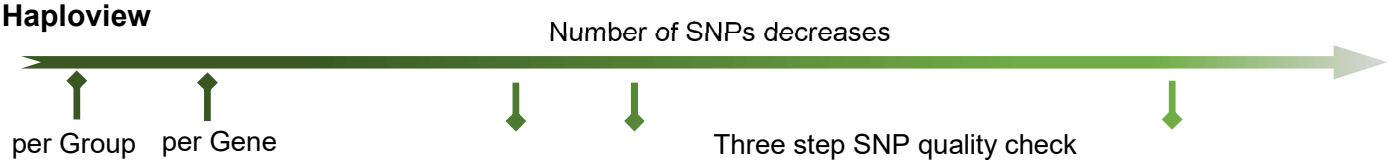
Minor allele carrier - Minimum of 10 samples either heterozygous or homozygous for the minor allele

Coloured/Highlighted SNPs were removed

This is the SNP Selection and QC for the African population group. The genes were arranged according to the chromosome number. The SNPs were arranged according to decreasing MAF. All the SNPs that had no observable minor alleles were removed. The x^2 (was calculated by the ObsHET and PredHET) and HWE served as mutation interference QC parameters (cut-off $p < 0.05$). The Minor allele carriers determines the minimum of 10 samples needed for sufficient association analysis not only for the combined homozygous minor and heterozygous genotypes but also the homozygous wild genotype.

Table 4.4 contains the 15 SNPs for the African Control group that exhibited stable population allele frequencies and had at least 10 minor allele carriers (heterozygous or homozygous for the minor allele) and were, therefore, suitable for use in further association analysis in PLINK and SPSS. The same applied for the Combined Controls group, Table 4.5 shows the 26 SNPs in the Combined Controls group that were selected for association analysis in PLINK.

Table 4.5 Combined Controls population group SNP selection and Quality Control within Haploview



Population	Gene	SNPs	x^2	HWE - *pval*	▼MAF▼	Sample Size	Minor allele carriers	Alleles
Combined	*CYP1B1*	rs162562	0.10	0.05				
		rs1056837	0.02					
		rs1056836	0.01					C:G
		rs10012	0.41	0.39	0.37	44	25	G:C
		rs2551188	0.37	0.32	0.35	46	25	C:T
		rs10916	0.22	0.09	0.31	46	21	A:C
		rs1800440	0.78	0.49	0.13	46	11	T:C
		rs4986889	0.97	1.00	0.03		3	
		rs4987134	1.00	1.00	0.02		2	
		rs4986888	1.00	1.00	0.01		1	
		rs57865060	1.00	1.00	0.01		1	
		The Rest	0.00					
	CYP3A4	rs4646437	0.35	0.00				
		rs3735451	0.30	0.00				
		rs2242480	0.47	0.00				
		rs2687116	0.93	0.01				
		rs2246709	0.93	1.00	0.21	48	18	A:G
		rs6956344	0.83	1.00	0.20	48	17	C:T
		rs4646440	0.83	1.00	0.04		4	
		rs35599367	0.76	1.00	0.03		3	
		rs28988600	0.76	1.00	0.02		2	
		rs28371757	0.42	1.00	0.01		1	
		rs4986907	0.52	1.00	0.01		1	
		rs34642455	0.96	1.00	0.01		1	
		rs28988574	0.41	1.00	0.01		1	
		The Rest	0.00					
	CYP1A1 / CYP1A2	rs2606345	0.02					
		rs2472304	0.25	0.13	0.39	48	27	G:A
		rs2470890	0.08	0.01				
		rs762551	0.68	0.80	0.32	48	27	A:C
		rs4646421	0.70	0.63	0.21	48	17	G:A
		rs17861157	0.85	1.00	0.08		8	
		rs2069526	0.66	0.42	0.07		6	
		rs17861094	0.61	0.31	0.06		5	
		rs4646427	0.61	0.31	0.06		5	
		rs4986883	0.55	0.21	0.05		4	
		rs17861084	0.96	1.00	0.04		4	
		rs1799814	0.98	1.00	0.03		3	
		rs3743484	0.98	1.00	0.02		2	
		rs4986884	1.00	1.00	0.01		1	
		rs4986881.1	1.00	1.00	0.01		1	

		rs41279188	1.00	1.00	0.01		1	
		rs1048943.3	1.00	1.00	0.01		1	
		rs17861152	1.00	1.00	0.01		1	
		rs45540640	1.00	1.00	0.01		1	
		rs56276455	1.00	1.00	0.01		1	
		The Rest	0.00					
	COMT	rs4633	0.43	0.35	0.50	48	34	C:T
		rs5993883	0.07	0.02				
		rs165599	0.16	0.07	0.49	48	32	G:A
		rs740603	0.36	0.28	0.48	47	32	G:A
		rs4680	0.56	0.53	0.47	48	33	G:A
		rs174696	0.47	0.40	0.44	48	31	T:C
		rs5746849	0.74	0.80	0.44	48	32	G:A
		rs933271	0.18	0.06	0.34	48	24	T:C
		rs2239393	0.59	0.65	0.33	48	28	A:G
		rs3819619	0.99	1.00	0.32	48	26	G:A
		rs6269	0.79	0.97	0.31	48	26	A:G
		rs4818	0.75	0.93	0.24	48	21	C:G
		rs4646316	0.84	0.92	0.23	48	19	C:T
		rs9332377	0.41	0.19	0.21	48	16	C:T
		rs737866	0.70	0.63	0.21	48	17	T:C
		rs737865	0.70	0.63	0.21	48	17	A:G
		rs4646312	0.70	0.83	0.20	48	18	T:C
		rs35788262	0.86	0.95	0.12	48	10	G:A
		rs769224	0.61	0.31	0.06		5	
		rs8192488	0.47	0.14	0.05		3	
		rs174699	0.98	1.00	0.03		3	
		rs165631	0.98	1.00	0.02		2	
		rs34068292	1.00	1.00	0.01		1	
		The Rest	0.00					

Legend

African - African Controls

CYP - Cytochrome P45

COMT – Catechol-O-methyltransferase

The Rest - The remaining SNPs

x^2 - Chi Square test

HWE - Hardy Weinberg equilibrium

MAF - Minor allele frequency

pval - p-value

Minor allele carrier - Minimum of 10 samples either heterozygous or homozygous for the minor allele

Coloured/Highlighted SNPs were removed

This is the SNP Selection and QC for the Combined Controls population group. The genes were arranged according to the chromosome number. The SNPs were arranged according to decreasing MAF. All the SNPs that had no observable minor alleles were removed. The x^2 (was calculated by the ObsHET and PredHET) and HWE served as mutation interference QC parameters (cut-off $p < 0.05$). The Minor allele carriers determines the minimum of 10 samples needed for sufficient association analysis not only for the combined homozygous minor and heterozygous genotypes but also the homozygous wild genotype.

All these selected SNPs from Tables 4.3-4.5 had a non-significant (x^2) value indicating there are no statistical differences between the ObsHET and PredHET groups (Curtis & Youngquist, 2013). Furthermore, all the non-significant HWE-*p*val obtained indicate that multiple factors such as biased mutations within the SNP-samples are not prevalent enough to influence downstream association analysis calculations (Ryckman & Williams, 2008). Also, all these SNPs had a sufficient sample size of 10 samples that at least carried a single minor allele. With these SNP selection steps in Haploview the requirements for objective 6 were reached.

After selecting SNPs, the chances for non-random inheritance patterns were also calculated in Haploview. This non-random linkage is better known as linkage disequilibrium (LD) on the basis of the SNPs being in proximity on the chromosome and with a low in between recombination rate to form a group of SNPs called haplotype/block (Gabriel *et al*., 2002). No reference SNPs were selected since the limited LD results that small sample sizes could produce caused more confusion than lessened workload.

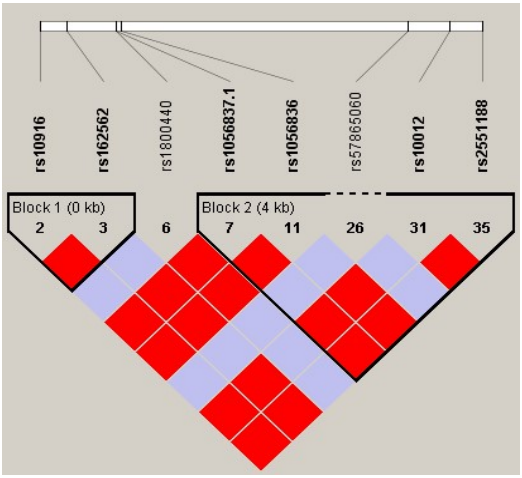
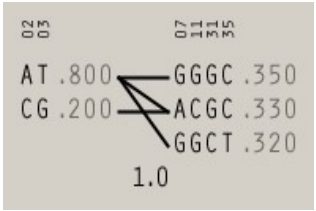
The LD-plots in Figure 4.3 illustrate all SNPs for a given gene that demonstrate minor allele frequencies per population group. The SNPs that were strongly, non-randomly linked are outlined by black triangular shapes. In addition, using the same calculations as LD, the Hendricks Multiallelic D' (Multiallelic D) was also determined which indicates the degree of linkage / recombination between haplotypes (Barrett *et al*., 2005). The Multiallelic D linkage strength is indicated by the weight of the lines that visually connect the haplotypes.

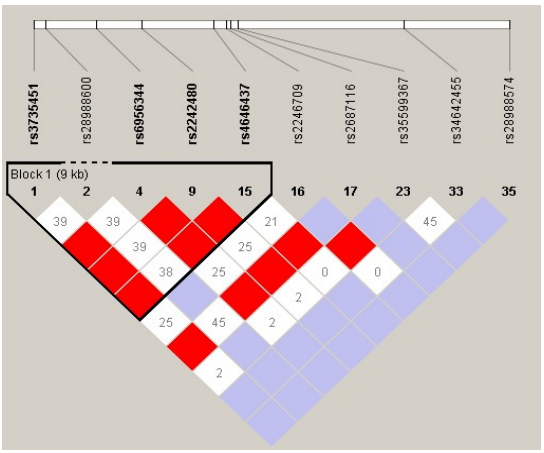
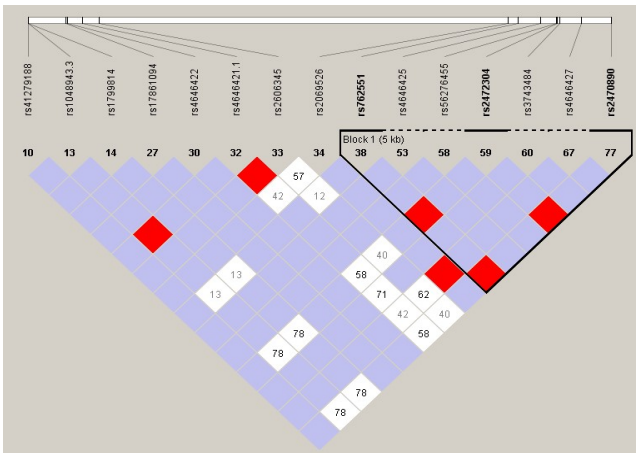
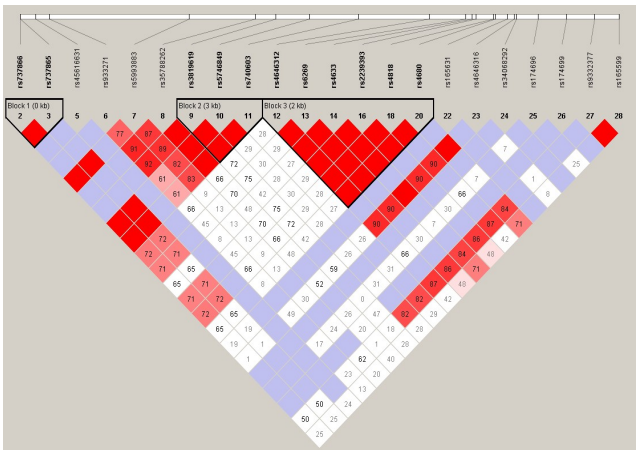
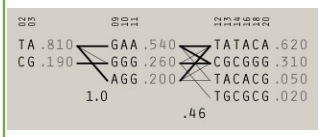
As can be seen from Figure 4.3, the SNPs rs10916 and rs162562; as well as rs1056837.1, rs1056836, rs10012 and rs2551188 of *CYP1B1* are in LD forming two haploblocks for the Caucasian population with a multiallelic D value above 0.95 indicating the haploblocks can be linked together (Gabriel *et al*., 2002), although only the latter haploblock contains some selected SNPs for further association analysis. The possible reasons for rs1800440 not being within the haplotype is either due to small sample size or that it is a recombination region. These linkages might indicate that certain SNPs needed to be inherited together in order to influence metabolic levels. Within the African population only the two selected SNPs (rs10012 and rs2551188) are in LD forming a single haploblock. For the Combined controls group, the SNPs rs10916 and rs162562; as well as 10012 and rs2551188 are in LD forming two haploblocks, both containing selected SNPs.

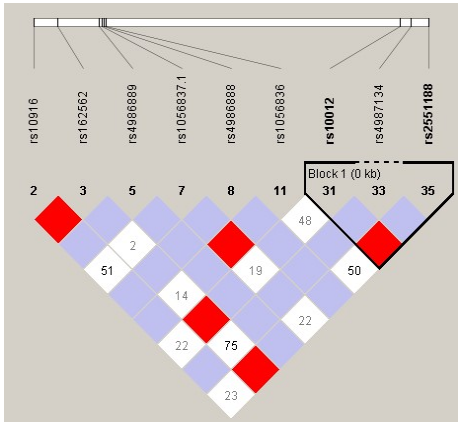
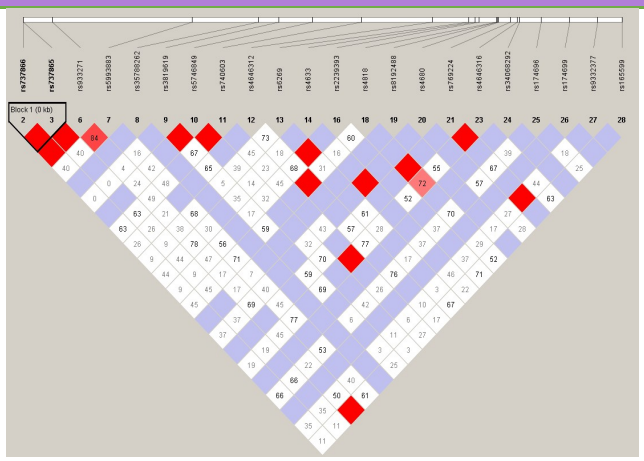
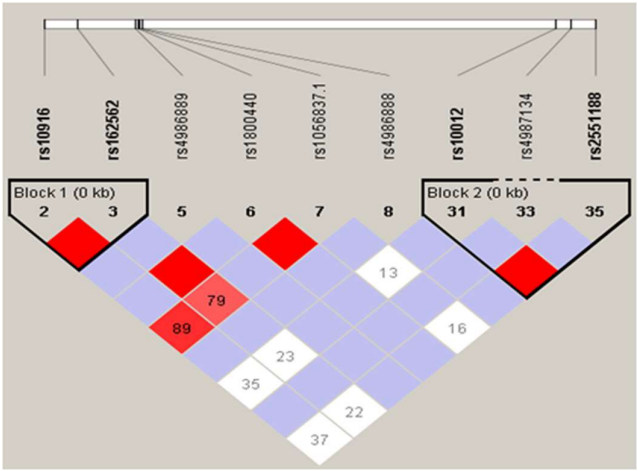
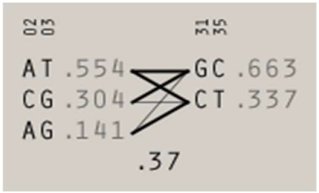
For the gene *CYP3A4* the SNPs rs3735451, rs6956344, 2242480, rs4646437 are in LD forming a haplotype for the Caucasian group, although none of these SNPs were selected for further association analysis. Within the African and Combined controls groups no LD could be calculated. Therefore, *CYP3A4* has no SNPs in any population in LD for the purposes of this study.

The SNPs rs762551, rs2472304 and rs2470890 of *CYP1A1/1A2* are in LD forming a single haploblock containing some selected SNPs in the Caucasian groups. Within the African population no LD of SNPs could be calculated. The SNPs rs17861094, rs4646421.1 and rs2069526 were in LD forming a single haploblock containing none of the selected SNPs.

For the gene *COMT* the SNPs rs737866 and rs737865; as well as rs3819619, rs57849 and rs740603, and lastly rs4646312, rs6269, rs4633, rs2239393, rs4818 and rs4680 are in LD forming three haploblocks with only the last haploblock containing selected SNPs, although haploblock one and two could be inherited together due to a very high multiallelic D value of above 0.95 (Gabriel *et al*., 2002). The SNPs rs737866 and rs737865 are the only SNPs in LD forming a haploblock in the African population that contains no selected SNPs. Within the Combined controls group the SNPs rs737866 and rs737865; as well as rs5746849 and rs740603; and lastly SNPs rs6269, rs4633 and rs2239393 are in LD forming three haploblocks of which all the SNPs were selected SNPs.

		Illustration Type	
Population Group	Gene	LD-plot	Hendricks Multiallelic D'
(a) Caucasian	*CYP1B1*		

	CYP3A4		
	CYP1A1 / CYP1A2		
	COMT		

(b) African	*CYP1B1*		
	CYP3A4	No haplotype / blocks	
	CYP1A1 / CYP1A2	No haplotype / blocks	
	COMT		
(c) Combined Controls	*CYP1B1*		

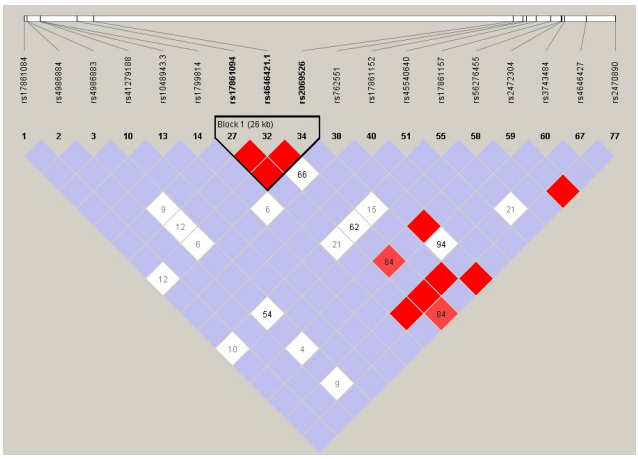
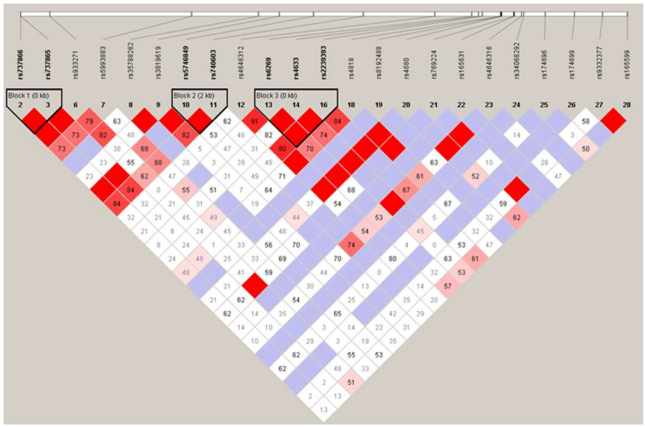
	CYP3A4	No haplotype / blocks	
	CYP1A1 / CYP1A2		
	COMT		

Figure 4.3 Haploview LD-plots and Haplotype Linkages for all groups and genes

Legend

LD-plot - Linkage Disequilibrium plot; African - African Controls; Caucasian - Caucasian Controls and COC users; COC - Combined oral contraceptive users (Caucasian); Combined - African and Caucasian Controls; *CYP* - Cytochrome P450; *COMT* – Catechol-O-methyltransferase

The graphs show the SNPs for the *CYP1B1, CYP3A4, CYP1A1/1A2* and *COMT* genes that were in linkage disequilibrium (LD-plots) as well as Hendricks Multiallelic D' values per population group. The Caucasian population group consisted of both Caucasian Controls and COC users. SNPs that were calculated to have linked frequencies were grouped together in haploblocks/types and are outlined on the LD-plots with black triangular shapes. The Hendricks Multiallelic D' demonstrates the degree of linkage between the haploblocks. The weight of the black line correlates with the degree of linkage.

In Figure 4.3 all three population groups demonstrated LD. The Combined Controls haplotypes contain less SNPs than the Caucasian group haplotypes. The African population group has

even less SNPs per haplotype as well as fewer pronounced haplotypes. This is due to SNP allele frequency being highly correlated in the homogeneity of the Caucasian population and is more distinct in the heterogeneous African populations (Shifman *et al.,* 2003). The Combined Controls demonstrated how mixed ethnicity influences haplotype formation.

4.4 Phenotypic data Quality Control

From here on the data were analysed in the four main population groups (as indicated in Table 4.6). As seen in Table 4.6, an average of 80% of the unaltered metabolites were considered non-normally distributed according to the Shapiro-Wilk skewness and kurtosis parameters. Shapiro-Wilk is an appropriate normality test since it is statistically powerful and is suggested for sample sizes below 50 (Razali & Wah, 2011). The skewness parameter calculates the symmetry of the distribution, and the kurtosis parameter is a measure of the distribution tail length or non-directly the distribution peak height (Razali & Wah, 2011). The phenotype data was log transformed, where after the outliers that were three standard deviations higher than the mean were removed in order to improve the fit and assumptions of association analysis. This method was used to conserve as much data as possible within the small sample size per group.

Table 4.6 All 4 groups Shapiro-Wilk Normalization

Population Groups	**Descriptive Statistics**	
	Amount of Metabolite Distributions that Deviated from Normal Distribution	
	Unaltered Metabolites	Transformed Metabolites with Outliers Removed
Caucasian Control	29 of 36	5 of 36
African Control	29 of 36	5 of 36
Combined Controls	32 of 36	8 of 36
COC	28 of 36	3 of 36

Legend

Combined - Caucasian and African Controls; COC - Combined oral contraceptive users (Caucasian)

This table lists the non-normal distributed metabolites in different population groups from unaltered to log10 transformed. There are 36 metabolites that are oestrogen metabolite, conjugates and adducts as well as redox balance metabolites. The non-normally distributed metabolites were still used for association analysis.

Table 4.6 above demonstrates how log transformation reduced the number of metabolites that deviated from a normal distribution to about 15% of the transformed metabolites. Normal distribution curve parameters adjusted for small sample sizes are according to Byrne (2016) considered to have a skewness between -2 and +2 and kurtosis between +7 and -7. The calculated Shapiro–Wilk skewness and kurtosis parameters can be seen per population group in Appendix Table S0.4-S0.7.

The non–normally distributed variables were still subjected to parametric association analysis, since the association software's (PLINK's) logistic command is flexible with non-normally distributed metabolites (Clarke *et al*., 2011) and the SPSS general linear model (GLM) is known to tolerate moderate deviations from normal distributions (Kirpich *et al*., 2018) fulfilling the requirements of objective 7.

4.5 PLINK and SPSS association analysis

The preliminary association analyses with initial processing of the selected SNPs and adjusted metabolite data were done using PLINK (Table 4.7). This is preliminary, given that PLINK can only indicate association and not which genotype of the SNP is associated with the metabolite. Therefore, this was followed by another association analysis in SPSS for the final QC and determination of more accurate significance along with the covariate and genotype consideration.

For clarity, PLINK applies a range of non-parametric tests to reveal allele variance within a population (Alhusain & Hafez, 2018). Therefore, non-normal distributed metabolites could be included in preliminary association and the association was done per population group. The resulting *p*-values were adjusted by the False Discovery Rate of Benjamini–Hochberg (FDR-BH) to avoid false positives resulting from multiple association calculations (Yekutieli, 2008). Since PLINK association analysis have little to no false positive results according to Kirpich *et al.* (2018), and a loss of power can occur in PLINK with association analyses in small samples sizes (Zhang *et al.,* 2002), metabolites with *p*-values up to 0.08 were included to avoid the exclusion of false negative results. Nevertheless, the FDR-BH was still the strongest factor applied during Quality Control of the association data.

As seen in the association results in Table 4.7, for the Caucasian Controls, PLINK identified 25 associations with mostly oestrogen-conjugates and also E_3, reactive oxygen species (ROS) and DNA-adducts. The 12 SNPs that associated are found in the phase I *CYP1B1* and *CYP1A2* genes and the phase II *COMT* gene.

The African Control group had 15 associations. The eight SNPs that associated with hydroxylated oestrogens, methoxy-oestrogens, E_2 conjugates, ROS and E_1 are found in the phase I *CYP3A4* and *CYP1A2* genes, as well as the phase II *COMT* gene.

The Combined Controls group demonstrated 55 associations with mostly pre-phase I hormones such as oestrone, testosterone and oestrogen conjugates as well as ROS, ferric reducing ability of plasma (FRAP) and glutathione-S-transferase (GSHt). The 22 SNPs that associated are from all five genes of interest. The last population (COC) had the least associations – only about 12. The associations were with various types of metabolites such as conjugates, two intermediate oestrogens as well as a DNA-oestrogen adduct. The 5 SNPs were form the *CYP1B1, CYP1A1* and *COMT* genes. These preliminary associations reduced the possible thousand associations to 107 associations for detailed processing thereby adequately achieving objective 8.

For further analyses in SPSS, all 29 SNPs and 23 metabolites that appear in Tables 4.7 were included. Association analyses were done on this data set using GLM along with covariates to take metabolite association adjustments as well as genotype sample size into consideration. The four chosen covariates for this study that is known to influence metabolic levels and consequently, association analysis as stated by Aschard *et al.* (2017) are BMI, age, menstrual phase and, specifically applied to the Combined Controls group, ethnicity (race).

Tables 4.8 list the accurate and significant associations of all the population groups and summarizes the results of the SPSS parametric test-GLM univariate model. The analysis was done 1) per population group, 2) per gene, 3) per appropriate PLINK association, and 4) per allele model, using the transformed metabolite data to calculate significance.

There are various allele models in literature as explained by Kirpich *et al*., (2018) and in this study the minor allele perspective model was used. The minor allele model further consists of three models with “M” as the Major allele and “m” as the minor allele as determined per study group in the previous Table 4.3-4.5. The **first** model is the Additive (co-dominant) model that compares all three genotypes namely the homozygous wild type (MM) genotype, heterozygous (Mm) genotype and homozygous for the minor allele (mm) genotype. This model enables a specific single genotype to be linked to a single metabolite. The **second** model is Dominant model consisting of the homozygous wild type (MM) genotype and both the other groups combined into one (Mm + mm) to determine if the minor allele has a dominant effect.

Table 4.7 PLINK All population groups Preliminary Association Analyses

Caucasian Controls

Significant

Genes	Metabolite	SNP	▲FDR -BH▲
COMT	E_3-16-glucu	rs4818	< 0.001
COMT	E_3-16-glucu	rs4646312	< 0.001
COMT	E_3-16-glucu	rs6269	0.001
COMT	E_3-16-glucu	rs2239393	0.001
COMT	E_3-16-glucu	rs4633	0.012
COMT	E_3-16-glucu	rs4680	0.012
COMT	E_3-16-glucu	rs165599	0.018
CYP1B1	E_2-3-glucu	rs2551188	0.03
CYP1B1	E_2-3-glucu	rs10012	0.03
CYP1B1	E_1-3-glucu	rs2551188	0.036
FCYP1B1	E_1-3-glucu	rs10012	0.036

Possible False Negatives

Genes	Metabolite	SNP	▲FDR -BH▲
CYP1A2	E_3	rs762551	0.052
CYP1A2	E_3	rs2470890	0.052
CYP1A2	E_3	rs2472304	0.052
COMT	E_3-3-S	rs4818	0.055
COMT	E_3-3-S	rs4646312	0.055
COMT	E_3-3-S	rs6269	0.055
COMT	E_3-3-S	rs2239393	0.055
COMT	E_3-3-S	rs4633	0.055
COMT	E_3-3-S	rs4680	0.055
COMT	E_3-3-S	rs165599	0.055
CYP1B1	D-4OHE_1-1N3A	rs2551188	0.068
CYP1B1	D-4OHE_1-1N3A	rs10012	0.068
CYP1B1	ROS	rs10012	0.079
CYP1B1	ROS	rs2551188	0.079

African Controls

Significant

Genes	Metabolite	SNP	▲FDR -BH▲
COMT	4-HE_1	rs4680	0.005
COMT	E_1	rs4818	0.019
COMT	4-HE_2	rs4680	0.033
COMT	4-HE_2	rs4633	0.033
COMT	ROS	rs4680	0.034
COMT	ROS	rs4633	0.034
COMT	ROS	rs740603	0.034
CYP1A2	E_2-3-glucu	rs762551	0.04
COMT	2-HE_1	rs4680	0.044
COMT	2-HE_1	rs4818	0.044
COMT	4-HE_2	rs740603	0.049

Possible False Negatives

Genes	Metabolite	SNP	▲FDR -BH▲
CYP1A2	4-MOE_1	rs762551	0.054
CYP3A4	16-KetoE_2	rs6956344	0.059
CYP3A4	ROS	rs2246709	0.064
CYP3A4	16-KetoE_2	rs2687116	0.08

Combined Controls

Significant

Genes	Metabolite	SNP	▲FDR -BH▲
CYP1A2	GSHt	rs2472304	< 0.001
COMT	FRAP	rs4680	< 0.001
COMT	FRAP	rs4633	< 0.001
COMT	FRAP	rs165599	< 0.001
CYP1A2	FRAP	rs2472304	< 0.001
COMT	GSHt	rs4680	< 0.001
COMT	GSHt	rs165599	< 0.001
CYP1A1	GSHt	rs4646421	< 0.001
CYP1B1	GSHt	rs1800440	< 0.001
COMT	FRAP	rs174696	< 0.001

Possible False Negatives

Genes	Metabolite	SNP	▲FDR -BH▲
CYP3A4	Testosterone	rs2246709	0.051
COMT	GSHt	rs9332377	0.051
COMT	GSHt	rs35788262	0.051
COMT	GSHt	rs4646316	0.072
COMT	GSHt	rs740603	0.072
COMT	GSHt	rs2239393	0.073
COMT	E_3-16-glucu	rs4646312	0.075
COMT	E_3-16-glucu	rs4818	0.075
COMT	E_3-16-glucu	rs737865	0.075
COMT	E_3-16-glucu	rs737866	0.075

COMT	GSHt	rs4633	0.001
COMT	GSHt	rs174696	0.001
COMT	FRAP	rs4646316	0.001
CYP3A4	GSHt	rs6956344	0.002
COMT	FRAP	rs5746849	0.002
COMT	ROS	rs4680	0.003
COMT	ROS	rs4633	0.004
CYP3A4	ROS	rs6956344	0.005
CYP1B1	ROS	rs1800440	0.005
COMT	GSHt	rs3819619	0.006
COMT	GSHt	rs5746849	0.007
CYP1A1	FRAP	rs4646421	0.008
COMT	FRAP	rs3819619	0.008
CYP3A4	ROS	rs2246709	0.010
CYP1A2	ROS	rs2472304	0.011
CYP3A4	FRAP	rs6956344	0.012
CYP1A1	D-4OHE$_2$-1N3A	rs4646421	0.020
COMT	ROS	rs4646316	0.020
COMT	ROS	rs3819619	0.020
COMT	Testosterone	rs737865	0.021
COMT	Testosterone	rs737866	0.021
COMT	FRAP	rs740603	0.024
CYP1A1	4-HE$_2$	rs4646421	0.030
CYP1B1	4-MOE$_1$	rs10916	0.030
COMT	ROS	rs5746849	0.034
COMT	ROS	rs740603	0.034
COMT	FRAP	rs2239393	0.037
COMT	FRAP	rs933271	0.037
COMT	FRAP	rs9332377	0.037
COMT	E$_1$	rs4818	0.038
COMT	Testosterone	rs4680	0.039
COMT	Testosterone	rs5746849	0.039
COMT	Testosterone	rs740603	0.039
CYP1A1	ROS	rs4646421	0.048
CYP3A4	E$_2$-17-S	rs2246709	0.048

COC

Significant

Genes	Metabolite	SNP	▲FDR -BH▲
CYP1B1	16-KetoE$_2$	rs1800440	0.003
COMT	E$_2$-3-glucu	rs933271	0.034

Possible False Negatives

Genes	Metabolite	SNP	▲FDR -BH▲
CYP1A1	2-MOE$_1$	rs2606345	0.058
COMT	E$_1$-3-S	rs933271	0.063
CYP1B1	D-4OHE1-1N7G	rs1800440	0.065
CYP1B1	E$_2$-3-S	rs10012	0.066
CYP1B1	E$_2$-3-S	rs2551188	0.066
CYP1B1	D-4OHE2-1N3A	rs10012	0.069
CYP1B1	D-4OHE2-1N3A	rs2551188	0.069
CYP1B1	4-HE$_1$	rs10012	0.071
CYP1B1	4-HE$_1$	rs2551188	0.071

	CYP1B1	E_3-16-glucu	rs1800440	0.076

Legend

Combined Controls - African and Caucasian Controls; COC - Combined oral contraceptive users (Caucasian);
UNADJ - Unadjusted Values; FDR-BH - False Discovery Rate (Benjamini-Hochberg) (adjusted);
CYP - Cytochrome P450; *COMT* – Catechol-O -methyltransferase; $4OHE_{1/2}$-N (G / A) - oestrogen-DNA adduct;
E_1 - Oestrone; E_2 - Oestradiol; E_3 – oestriol; ROS - Reactive oxygen species; S – sulphate; glucu – glucuronide;
H – hydroxy; MO - Methoxy

Metabolite and SNP association are listed in this Table. The set of associations were arranged according to populations groups. The individual associations were arranged according to the increasing FDR-BH *p*-value. The FDR-BH *p*-value is a parameter against false positive associations that emerged from multiple statistic testing. Only the significant associations ($p < 0.08$) with sufficient sample sizes are displayed. The *p*-value from 0.05 to 0.08 is listed as possible false negative results. These associations were calculated with covariates and with post hoc comparisons.

The **third** model is the Recessive model where the homozygous minor (mm) genotype is compared to the homozygous wild type genotype and heterozygous genotype combined (MM + Mm) to determine if the minor allele has a recessive effect (Kirpich *et al*., 2018). Due to the small study sample size, some groupings did not have enough samples for meaningful association analyses. These are indicated with purple highlights in the “Sample Size (N)” rows in Table 4.8

The QC parameters for significant association is a *p*-value below 0.08 as explained for the PLINK results and sufficient statistical power of above 70% for small sample sizes (Button *et al.,* 2013). Moreover, the significant association genotypes should contain a sufficient number of samples (≥10 per group). The statistical power compares systematic with unsystematic variance to create a model of error and can moderately tolerate deviation from normal distributions (Field, 2013). A power below 80% could indicate false positive results and consequently not reflect the true effect (Button *et al.,* 2013). However, since small sample sizes are known to have definitive low power, the threshold for this study was changed to 70% and associations with power above 50% were flagged as a possible association. The statistical power was calculated with SPSS’s partial correlation method that is determined by effect size (value derived from GLM partial eta squared, η^2). The green highlighted *p*-values in all the *p*-value columns (Tables 4.8) indicate significant associations within the respective genotypes.

All of the association shown in Table 4.8 on the landscape pages 79 to 82, have a *p*-value below 0.08 and sufficient power above 70% except for one association that is in the COC group, namely rs10012 SNP which associated with 4-hydroxyoestrone had low power of 57%, as indicated with the dark square on Table 4.8.

In the Caucasian Control population group 15 of 25 associations were identified as significant in SPSS which is a large number of associations presented in Table 4.8. For all of these associations the SNP appeared to have a dominant minor allele. The African Control population group had five SPSS significant associations out of the 15 significant PLINK

associations. All five of these associations demonstrated that the minor allele is dominant. In the Combined Controls group only five of 55 associations were significant in SPSS. This is mostly due to the significant covariation in ethnicity [which will be discussed further on (Table 4.9)]. Every association within this population group identified the minor allele as dominant. Lastly, for the COC population group, four of 12 minor allele as dominant associations were significant in SPSS. The one association, *CYP1B1* rs10012 with ROS, had a low Power of 57%, indicating only a possible significant association.

These collectively 29 associations form the main part of objective 9 in attaining accurate SNP-metabolite associations per population group, although in order to conclude the main results of aim two, additional results are given in the form of mentioning an alternative method to support the chosen association analysis, the influence of covariates, a summary of the SNP-metabolite selection and a *post-hoc* test for all interested before continuing to the discussion.

As an alternative approach, another type of test statistic, namely the non-parametric Mann-Whitney U test that does not depend on metabolite normal distributions was used to compare types of associations (Chakraborti & Wiel, 2008). The Mann-Whitney U test was done on randomly selected data points to include one significant and one non–significant association, and thereafter one normally and one non–normally distributed metabolite per population group. The *p*-values obtained from this analysis were compared to the GLM *p*-values and are given in appendix Table S0.8. Even though the non-parametric test was slightly more accurate (since the p-values were slightly more significant due to circumvention of slight skewness) than GLM since the test is not influenced by metabolite distributions, it did not include covariate testing (Chakraborti & Wiel, 2008). GLM, therefore, proved to be sufficiently accurate and robust even for the non-normally distributed metabolites. This confirms the suitability of GLM association analysis of small sample sizes and non-normally distributed variables.

Table 4.8 Significant Associations for all population groups according to SPSS GLM Model

Group	Gene										
Caucasian Control	**CYP1B1**										
Metabolite	SNP	Allele Models from the Minor Allele perspective **(M/m - Major/minor Allele)**									
	rs2551188	Additive (MM + Mm + mm)				Dominant (MM + Mm/mm)			Recessive (MM/Mm + mm)			Power (%)
	Sample Size (N)	13 (MM)	10 (Mm)	2 (mm)	*p*-val	13 (MM)	12 (Mm + mm)	*p*-val	23 (MM + Mm)	2 (mm)	*p*-val	
ROS		1.88 ± 0.06	1.82 ± 0.06	1.84 ± 0.1	0.116	1.88 ± 0.06	1.82 ± 0.06	0.043	1.86 ± 0.07	1.84 ± 0.1	0.920	76.3
	rs10012											
	Sample Size (N)	12 (MM)	10 (Mm)	2 (mm)	*p*-val	12 (MM)	12 (Mm + mm)	*p*-val	22 (MM + Mm)	2 (mm)	*p*-val	
ROS		1.89 ± 0.06	1.82 ± 0.06	1.84 ± 0.1	0.067	1.89 ± 0.06	1.82 ± 0.06	0.022	1.86 ± 0.07	1.84 ± 0.1	0.886	85
	CYP1A2											
	SNP	Allele Models from the Minor Allele perspective **(M/m - Major/minor Allele)**										
	rs2470890 / rs2472304											
		Additive (MM + Mm + mm)				Dominant (MM + Mm/mm)			Recessive (MM/Mm + mm)			Power (%)
	Sample Size (N)	10 (MM)	12 (Mm)	3 (mm)	*p*-val	10 (MM)	15 (Mm + mm)	*p*-val	22 (MM + Mm)	3 (mm)	*p*-val	
E_3		0.36 ± 0.24	0.69 ± 0.41	0.78 ± 0.66	0.151	0.36 ± 0.24	0.71 ± 0.44	0.054	0.54 ± 0.38	0.78 ± 0.66	0.380	72.1
	COMT											
	SNP	Allele Models from the Minor Allele perspective **(M/m - Major/minor Allele)**										
	rs4818 / rs4646312											
		Additive (MM + Mm + mm)				Dominant (MM + Mm/mm)			Recessive (MM/Mm + mm)			Power (%)
	Sample Size (N)	14 (MM)	11 (Mm)	0 (mm)	*p*-val	14 (MM)	11 (Mm + mm)	*p*-val	25 (MM + Mm)	0 (mm)	*p*-val	
E_3-16-glucu		2.61 ± 0.46	1.77 ± 0.34	-	< 0.001	2.61 ± 0.46	1.77 ± 0.34	< 0.001	2.24 ± 0.59	-	-	99.7
E_3-3-S		1.51 ± 0.76	0.86 ± 0.41	-	0.041	1.51 ± 0.76	0.86 ± 0.41	0.041	1.23 ± 0.7	-	-	76.8
	rs6269 / rs2239393											
	Sample Size (N)	14 (MM)	10 (Mm)	1 (mm)	*p*-val	14 (MM)	11 (Mm + mm)	*p*-val	24 (MM + Mm)	1 (mm)	*p*-val	
E_3-16-glucu		2.61 ± 0.46	1.74 ± 0.34	2.08	< 0.001	2.61 ± 0.46	1.77 ± 0.34	< 0.001	2.25 ± 0.6	2.08	0.577	99.7
E_3-3-S		1.51 ± 0.76	0.86 ± 0.43	0.86	0.128	1.51 ± 0.76	0.86 ± 0.41	0.041	1.24 ± 0.71	0.86	0.413	76.8
	rs4633 /											

	rs4680											
	Sample Size (N)	12 (MM)	11 (Mm)	2 (mm)	*p*-val	12 (MM)	13 (Mm + mm)	*p*-val	23 (MM + Mm)	2 (mm)	*p*-val	
E_3-3-S		1.58 ± 0.8	0.88 ± 0.42	1.02 ± 0.22	0.120	1.58 ± 0.8	0.9 ± 0.39	0.038	1.24 ± 0.73	1.02 ± 0.22	0.633	78.2
	rs165599											
	Sample Size (N)	12 (MM)	13 (Mm)	0 (mm)	*p*-val	12 (MM)	13 (Mm + mm)	*p*-val	25 (MM + Mm)	0 (mm)	*p*-val	
E_3-16-glucu		2.53 ± 0.54	1.97 ± 0.5	-	0.052	2.53 ± 0.54	1.97 ± 0.5	0.052	2.24 ± 0.59	-	-	72.9
Group	**Gene**											
African Control	***CYP3A4***											
Metabolite	SNP	Allele Models from the Minor Allele perspective **(M/m - Major/minor Allele)**										
	rs2246709	Additive (MM + Mm + mm)				Dominant (MM + Mm/mm)			Recessive (MM/Mm + mm)			Power (%)
	Sample Size (N)	11 (MM)	12 (Mm)	0 (mm)	*p*-val	11 (MM)	12 (Mm + mm)	*p*-val	23 (MM + Mm)	0 (mm)	*p*-val	
ROS		1.92 ± 0.08	1.99 ± 0.05	-	0.016	1.92 ± 0.08	1.99 ± 0.05	0.016	1.95 ± 0.07	-	-	88,4
	COMT											
	SNP	Allele Models from the Minor Allele perspective **(M/m - Major/minor Allele)**										
	rs4680	Additive (MM + Mm + mm)				Dominant (MM + Mm/mm)			Recessive (MM/Mm + mm)			Power (%)
	Sample Size (N)	13 (MM)	10 (Mm)	0 (mm)	*p*-val	13 (MM)	10 (Mm + mm)	*p*-val	23 (MM + Mm)	0 (mm)	*p*-val	
4-HE_1		1.61 ± 0.67	0.66 ± 0.42	-	0.003	1.61 ± 0.67	0.66 ± 0.42	0.003	1.19 ± 0.74	-	-	96.5
ROS		1.99 ± 0.05	1.91 ± 0.08	-	0.014	1.99 ± 0.05	1.91 ± 0.08	0.014	1.95 ± 0.07	-	-	89.4
	Sample Size (N)	12 (MM)	10 (Mm)	0 (mm)	*p*-val	12 (MM)	10 (Mm + mm)	*p*-val	22 (MM + Mm)	0 (mm)	*p*-val	
2- HE_1		1.74 ± 0.77	0.89 ± 0.52	-	0.016	1.74 ± 0.77	0.89 ± 0.52	0.016	1.35 ± 0.79	-	-	88.2
	rs4633	12 (MM)	9 (Mm)	2 (mm)	*p*-val	12 (MM)	11 (Mm + mm)	*p*-val	21 (MM + Mm)	2 (mm)	*p*-val	
ROS		1.98 ± 0.05	1.94 ± 0.08	1.84 ± 0.08	0.019	1.98 ± 0.05	1.92 ± 0.09	0.049	1.96 ± 0.06	1.84 ± 0.08	0.015	74.4
Group	**Gene**											
Combined Controls	***CYP3A4***											
Metabolite	SNP	Allele Models from the Minor Allele perspective **(M/m - Major/minor Allele)**										

	rs2246709	Additive (MM + Mm + mm)				Dominant (MM + Mm/mm)			Recessive (MM/Mm + mm)			Power
	Sample Size (N)	29 (MM)	16 (Mm)	2 (mm)	*p*-val	29 (MM)	18 (Mm + mm)	*p*-val	45 (MM + Mm)	2 (mm)	*p*-val	
E_2-17-S		0.5 ± 0.38	0.78 ± 0.34	0.74 ± 0.1	0.015	0.5 ± 0.38	0.77 ± 0.32	0.005	0.6 ± 0.39	0.74 ± 0.1	0.86	95.7
Testosterone		0.38 ± 0.29	0.66 ± 0.42	0.55 ± 0.25	0.062	0.38 ± 0.29	0.65 ± 0.4	0.018	0.48 ± 0.36	0.55 ± 0.25	0.344	87.2
	COMT											
	SNP	Allele Models from the Minor Allele perspective **(M/m - Major/minor Allele)**										
	rs4818	Additive (MM + Mm + mm)				Dominant (MM + Mm/mm)			Recessive (MM/Mm + mm)			Power
	Sample Size (N)	27 (MM)	19 (Mm)	2 (mm)	*p*-val	27 (MM)	21 (Mm + mm)	*p*-val	46 (MM + Mm)	2 (mm)	*p*-val	
E_1		1.61 ± 0.52	2.04 ± 0.69	2.75 ± 0.98	0.032	1.61 ± 0.52	2.11 ± 0.72	0.021	1.79 ± 0.63	2.75 ± 0.98	0.082	85.7
E_3-16-glucu		2.36 ± 0.62	2.08 ± 0.52	1.31 ± 0.84	0.029	2.36 ± 0.62	2 ± 0.58	0.066	2.24 ± 0.59	1.31 ± 0.84	0.023	100
	rs4646312											
	Sample Size (N)	30 (MM)	17 (Mm)	1 (mm)	*p*-val	30 (MM)	18 (Mm + mm)	*p*-val	47 (MM + Mm)	1 (mm)	*p*-val	
E_3-16-glucu		2.4 ± 0.61	1.89 ± 0.52	1.9 ± 0.52	0.030	2.4 ± 0.61	1.89 ± 0.51	0.008	2.21 ± 0.63	1.9 ± 0.52	0.515	93.8
Group	**Gene**											
COC	***CYP1B1***											
Metabolite	SNP	Allele Models from the Minor Allele perspective **(M/m - Major/minor Allele)**										
	rs1800440	Additive (MM + Mm + mm)				Dominant (MM + Mm/mm)			Recessive (MM/Mm + mm)			Power (%)
	Sample Size (N)	13 (MM)	12 (Mm)	0 (mm)	*p*-val	13 (MM)	12 (Mm + mm)	*p*-val	25 (MM + Mm)	0 (mm)	*p*-val	
16-KetoE_2		1.11 ± 0.51	0.44 ± 0.39	-	0.003	1.11 ± 0.51	0.44 ± 0.39	0.003	0.79 ± 0.56	-	-	96.4
4OHE1-1N7G		1.74 ± 1.15	0.76 ± 0.78	-	0.051	1.74 ± 1.15	0.76 ± 0.78	0.051	1.27 ± 1.09	-	-	73.2
E_3-16-glucu		2.1 ± 0.38	1.7 ± 0.47	-	0.045	2.1 ± 0.38	1.7 ± 0.47	0.045	1.91 ± 0.46	-	-	75.4
	rs10012											
	Sample Size (N)	12 (MM)	8 (Mm)	5 (mm)	*p*-val	12 (MM)	13 (Mm + mm)	*p*-val	20 (MM + Mm)	5 (mm)	*p*-val	
4-HE_1		0.79 ± 0.36	0.96 ± 0.65	1.41 ± 0.75	0.069	0.79 ± 0.36	1.13 ± 0.7	0.063	0.86 ± 0.48	1.41 ± 0.75	0.037	57.1

Legend

CYP - Cytochrome P450; *COMT* - Catechol-O-methyltransferase; F - F score / statistic; ROS - Reactive oxygen species;

Combined Controls - African and Caucasian Controls; 4OHE1/2-N (G / A) - oestrogen-DNA adduct; E_1 - Oestrone; E_2 - Oestradiol; E_3 – Oestriol; S – sulphate; glucu – glucuronide; H – hydroxy; MO – Methoxy; COC - Combined oral contraceptive users (Caucasian)

Some significant associations are listed within this table. The associations are listed per population group, per gene and then per SNP. The transformed metabolite data was used to calculate the table mean, standard deviations and other values. Each SNP is sorted into three genotype groups namely the Additive, Dominant and Recessive models. The major and minor allele was determined by Ensembl All population group. The Additive model has all three genotypes. The Dominant model is the homozygous minor genotype and the combined other genotypes. The Recessive model is the homozygous wild genotype and the combined other genotypes. The sample sizes highlighted in purple were too small to do association analysis with. The green highlighted p-values are the significant associations. Covariates did not affect these p-values. The power and F-statistic was calculated from the significant association. The Power is the percentage to detect a significant association and the F statistic (higher value) is interpreted with the *p*-value (< 0.008).

Inclusion of covariates [specifically BMI, age, menstrual phase and (when applicable) ethnicity (race)] in association analyses raises the power of detecting accurate and significant association since the covariates could significantly influence p-values (Aschard *et al.,* 2017). All significant covariate associations are listed as categorical or continuous covariates in Table 4.9. The categorical covariates underwent SPSS GLM univariate analysis with the metabolites per population group. The continuous covariates were only detected in specific associations within a specific population group. Ethnicity is the most relevant covariate of the Combined Controls group and had 13 significant associations. These associations consisted of four metabolites [namely ROS, ferric reducing ability of plasma (FRAP), total red blood cellglutathione (GSHt) and Testosterone] which significantly differed between the African Control and Caucasian Control groups as well as one specific association between a DNA-adduct and the *CYP1A1* SNP (rs4646421) which differed. BMI as a continuous covariate had six significant associations, four in the African population and two in the Combined Controls, indicating that BMI could significantly contribute to the metabolite changes detected in the association analysis. The Age continuous covariate had four associations of which two were of importance, two in the Caucasian Control group with the oestriol-conjugates metabolite and the other two in the COC group with DNA-adducts. These results fully conclude objective 9.

Table 4.9 SPSS all population groups significant covariate association per covariate

Covariates		
Categorical covariates		
Ethnicity		
Combined Controls		
Metabolites		***p*-value**
GSHt		< 0.001
ROS		< 0.001
FRAP		< 0.001
Testosterone		0.031
	CYP1A1	***p*-value**
Metabolite	SNP	Covariate Association
4OHE$_2$-1N3A	Drs4646421	0.049
Menstrual Phase		
None		
Continuous covariates		
BMI		
African	**CYP3A4**	***p*-value**
Metabolite	SNP	Covariate Association
16-KetoE$_2$	Drs6956344	0.010
	COMT	
4-HE$_2$	Drs4680	0.046
4-HE$_2$	Rrs4633	0.034
ROS	Drs740603	0.050
Combined Controls	**CYP1B1**	***p*-value**
Metabolite	SNP	Covariate Association
4 -MOE$_1$	Drs10916	0.010
	CYP1A1	
4-HE$_2$	Drs4646421	0.026
Age		
Caucasian Control	**COMT**	***p*-value**
Metabolite	SNP	Covariate Association
E$_3$-16-glucu	Drs4633	0.058
E$_3$-16-glucu	Rrs4680	0.058
COC	**COMT**	***p*-value**
Metabolite	SNP	Covariate Association
4OHE$_2$-1N3A	Drs10012	0.002
4OHE$_2$-1N3A	R2551188	0.003

Legend

Combined Controls - African and Caucasian Controls; COC - Combined oral contraceptive users (Caucasian); 4OHE$_{1/2}$ - N (G / A) - oestrogen DNA adduct; E$_1$ - Oestrone; E$_2$ - Oestradiol; E$_3$ – Oestriol; ROS - Reactive oxygen species;

FRAP - Ferric reducing ability of plasma; GSHt – glutathione-S-transferase; glucu – glucuronide; H – hydroxy; MO – Methoxy; Testos – Testosterone; D-SNP – Dominant model; R-SNP – Recessive model

All covariate associations are listed within this table sorted by covariation groups. Each covariate group is arranged according to population group and then per gene. The transformed metabolite values were used to calculate these *p*-values. The capital D and R before every SNP ID is to refer to the dominant or recessive model from where the *p*-values associations were from. Only significant covariate associations are listed. All purple highlighted covariate *p*-values show which SNP-metabolite association were influenced by the covariate.

Table 4.10 summarizes the SNP and metabolite selection and association analyses process. The number of SNPs included at every step are indicated under each population group. This is shown per population group and not per gene. To summarize the process: all the SNPs listed on the GSA v2.0 annotated list were acquired from Genomestudio and the first, most drastic SNP selection QC step occurred in Haploview. The SNPs that passed this QC step were imported into PLINK together with the phenotype data that consisted of 36 metabolite levels for preliminary association analyses / from this, 107 associations emerged. The 29 SNPs and 22 metabolites that showed 107 associations in PLINK were then finally subjected to association analyses in SPSS where a total of 29 SNP-metabolite associations between 13 SNPs and 11 metabolites from the different population groups were identified from this study.

To determine which of the three genotypes in the Additive model were significantly different, a *Post hoc* test was done after association analysis in SPSS (Field, 2013). The *Post hoc* test is limited to Additive models with at least two samples for each of the three genotypes for a single SNP. *Post hoc* testing is not necessary for the recessive and dominant models since significant differences between only two groupings / genotypes are indicated by GLM *p*-value.

Table 4.10 Summary of SNP Selection of all the Genes

Selection Step	Caucasian Control	African Control	Combined Controls	COC	Total metabolites	Total associations
GSA v2.0	179	179	179	179		
Haploview	17	15	26	11	36	
PLINK	12	8	22	5	22	107
SPSS	**11**	**3**	**3**	**2**	**11**	**29**

Legend

GSA – Global Screening Array; Combined Controls - African and Caucasian Controls; COC - Combined oral contraceptive users (Caucasian)

This table demonstrates a summary of the flow of SNP and metabolite selection as well as association analysis. All SNPs were extracted from Genome Studio2.0. thereafter SNPs were selected with Haploview. PLINK preliminary selected SNPs and metabolites by association analysis. Lastly in SPSS the covariates and genotype applied specialised association analysis.

This study used Games Howell (G-H) *Post hoc* tests since it is statistically powerful and usually suggested for small sample size studies (Field, 2013). The *Post hoc* test results in Table 4.11

were calculated per population group, per gene, and then per significant association for the Additive model in Table 4.8.

Only five significant associations from Caucasian Control and three significant associations from Combined Controls study groups were revealed. These are indicated by green highlighted *p*-value in Table 4.11. There were no significant differences for the African controls. The significant associations for the Caucasian Control group were between *CYP1B1* rs10012 and ROS with the (homozygous) wild type differing from the heterozygous, between, both the *CYP1A2* rs2470890 and rs2472304 SNPs and oestriol with the heterozygous and (homozygous) minor groups differing significantly and sharing identical *p*-values, and between the *COMT* SNPs rs4633 and rs4680 and oestriol-3-sulphate with the (homozygous) wild type differing from the heterozygous, and the heterozygous differing from the (homozygous) minor genotype, respectively.

For the Combined Controls group there were 3 significant associations: *CYP3A4* rs2246709 associated with testosterone and oestradiol-17-sulphate with the (homozygous) wild type differing significantly from the heterozygous in both cases, while *COMT* rs4818 associated with oestrone with significant difference between the (homozygous) wild type and the heterozygous. Comparing the data of Tables 4.8 with that of Table 4.11 demonstrate how combining genotypes into either a Dominant or Recessive allele model can increase statistical power to detect significant associations that would have been otherwise missed – especially in studies with small sample sizes. Many SNP-metabolite associations only became apparent when the genotypes were specifically grouped together, according to a Dominant or Recessive allele model, while they were not significant when the Additive model was applied (Table 4.10-4.11). The additional detail provided by Tables 4.9-4.11 of the main results in Table 4.8 is used to fully achieve aim 2.

Table 4.11 SPSS Post hoc test for each Additive model SNP and Metabolite per Population Group

Group	Gene								
Caucasian Control	***CYP1B1***								
Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.	Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.
ROS	**rs2551188**	Wild Type	Hetero	0.147	ROS	**rs10012**	Wild Type	Hetero	0.025
			Minor	0.998				Minor	0.807
		Hetero	Minor	0.154			Hetero	Minor	0.942
	CYP1A2								
Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.	Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.
E_3	**rs2470890**	Wild Type	Hetero	0.969	E_3	**rs2472304**	Wild Type	Hetero	0.969
			Minor	0.595				Minor	0.595
		Hetero	Minor	0.071			Hetero	Minor	0.071
	COMT								
Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.	Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.
E_3-3-S	**rs4633**	Wild Type	Hetero	0.044	E_3-3-S	**rs4680**	Wild Type	Hetero	0.788
			Minor	0.176				Minor	0.176
		Hetero	Minor	0.288			Hetero	Minor	0.044
African Control	***COMT***								
Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.					
ROS	**rs4633**	Wild Type	Hetero	0.457					
			Minor	0.353					
		Hetero	Minor	0.435					
Combined Controls	***CYP3A4***								
Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.	Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.
E_2-17-S	**rs2246709**	Wild Type	Hetero	0.041	Testos.	**rs2246709**	Wild Type	Hetero	0.066
			Minor	0.165				Minor	0.706
		Hetero	Minor	0.927			Hetero	Minor	0.876
	COMT								
Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.	Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.
E_1	**rs4818**	Wild Type	Hetero	0.066	E_3-16-glucu	**rs4818**	Wild Type	Hetero	0.233
			Minor	0.500				Minor	0.466
		Hetero	Minor	0.687			Hetero	Minor	0.590
COC	***CYP1B1***								
Metabolite	SNP	Genotype (1)	Genotype (2)	Signif.					
4-HE_1	**rs10012**	Wild Type	Hetero	0.776					

		Minor	0.275	
	Hetero	Minor	0.537	

Legend

Combined Controls - African and Caucasian Controls; COC - Combined oral contraceptive users (Caucasian); *CYP* - Cytochrome P450; 4OHE1/2-N (G / A) - oestrogen DNA adduct; E_1 - Oestrone; E_2 - Oestradiol; E_3 – Oestriol; S – sulphate; glucu – glucuronide; H – hydroxy; MO – Methoxy ROS - Reactive oxygen species, Signif. - Significant; Testos – Testosterone; Hetero – Heterozygous; Minor – homozygous minor genotype

This table compares SNP genotype frequencies that are linked to a specific metabolite. The post hoc test is done per population group and then per gene. The transformed metabolite data is used to calculate the table values. The green highlighted *p*-values demonstrates significant differences between genotypes. The significant associations that are not shown in this table only had two genotypes to compare instead of the three genotypes that are demonstrated by these SNP associations.

CHAPTER 5 – DISCUSSION

There are two primary aims (aim one and two) and another aim that is secondary (aim three). The first aims of selecting SNPs for association analysis consisted of six objectives as summarized per sentence in this section. The first objective was achieved by identifying four study groups consisting of different configurations of the 74 samples as listed in the methods chapter in Table 3.1. The five main genes (i.e., *CYP1A1*, *CYP1A2*, *CYP3A4*, *CYP1B1* and *COMT*) were thereafter identified. The gDNA of 74 samples were successfully isolated and determined to be high quality through spectrometry, fluorometry and electrophoresis-imaging measurements (Table 4.1 and Figure 4.1). All the samples sent to CPGR successfully generated data of numerous SNPs (±665 000) per sample. Of the large number of SNPs, all the SNPs forming part of the five main genes (179 SNPs) were selected and the quality of the SNP genotyping was supported by sufficient allele frequency calls (>98%), GenCall, Gentrain and Cluster Separation Scores measurements (Table 4.2). SNPs that suited the minimum requirements for association analysis such as x^2, HWE and minimum samples that carried minor allele were selected from Haploview and resulted in different results per study group (Table 4.3-4.5). All these objectives achieved aim one in selecting SNPs per study group for association analysis purposes.

The second aim of associating SNPs with metabolites per study group consisted of three objectives as described per sentence within this section. The metabolites that were used for association analysis were log transformed into acceptable normalization distributions according to the Shapiro-Wilk test as well as parameters determined by Byrne (2016) for small sample sizes and with tolerable deviation of both the downstream preliminary and detailed associations analysis. Preliminary association analysis done by PLINK further selected SNPs and metabolites and successfully highlighted specific SNP-metabolite associations per study group. The most important objective was the final, detailed 29 associations analysis generated by SPSS-GLM method (Table 4.8). Each of these 29 associations are further discussed in this chapter since the literature-sourced effect of the SNP on the enzyme activity or expression could explain the differences measured in metabolite levels.

This discussion expands on the results of objective 9, the main objective of aim two, and objective 10 (the last objective of this study) of aim three will compare the results between specific study groups.

5.1 SNPs and Association Discussion

This study focusses primarily on association analysis of specific SNPs with specific metabolites and used literature to clarify SNP effect on enzyme catalytic activity, rate of expression and metabolite pathways. The 29 associations are discussed per population group. It should be noted that the validity and power of SNP-metabolite associations should be applied to every association discussed within this document.

This study focused on SNP association with small metabolic intrapopulation differences within healthy women and not on vague association to any specific disease further explaining genetic factors influence on metabolite levels. These SNPs, similar to the changed metabolite levels, are possible biomarkers for risk in cancer formation especially breast cancer (BCa) and these SNPs that cause metabolic changes are important to study within healthy (no measurable cancerous tissue) individuals (Frazer, 2009). Although this study contains small sample sizes, power was added to the analyses by ensuring a minimum of 10 samples within any group at any given time (e.g., 10 of 23 samples of our smallest study group) (Hertzog, 2008, de Winter 2009).

On the other hand, SNP-metabolite associations could reflect on the SNP effect on metabolite levels but does not necessarily define the direct biological effect, especially when the SNP is in LD with other SNPs and could be an effect of the combination of SNPs, or another SNP with the primary effect (Frayling, 2014). From another perspective, the absence of a cancer linked SNP does not assure that the disease will never form since genetics are only one of the factors involved in disease formation (Cavalieri & Rogan, 2014; Thomas *et al*., 2011). More factors can influence the protein expression and should be kept in mind when interpreting, such as the genes that encode specific transcription, translation and transfer proteins as well as accurate cofactors and dietary influences (Thomas *et al*., 2011). Quite a few SNPs' functions are based on other studies' hypotheses and has not yet been curated in enzyme studies (Thomas *et al*., 2011). Lastly SNPs are difficult to interpret due to differences in nomenclature, ambiguously and vaguely written association results along with the differences in metabolite measurement methods (Thomas *et al*., 2011).

Although all these limitations exist for association analysis, a proper attempt to describe the study's results could be made due to high quality control, high statistical power confirmation and the numerous external factor control. Within this discussion chapter most metabolites will be written in full to assist in comprehending the complex nature of the effect of association analyses.

5.1.1 Caucasian Control population

Within the Caucasian Control population group 15 out of 29 associations were found to be significant and included 11 SNPs (rs2551188 and rs10012 in *CYP1B1*, rs2470890 and rs2472304 in *CYP1A2* and rs4646312, rs4818, rs6269, rs165599, rs2239393, rs4680 and rs4633 in *COMT*) and with four metabolites (ROS, oestriol, oestriol-3-sulphate and oestriol-16-glucuronide) (Table 4.8). The first two associations were *CYP1B1* SNPs that associated with slightly decreased ROS. These SNPs form a haplotype where the minor allele is dominant in both rs2551188 (C > T) and rs10012 (G > C) according to Figure 4.3 (a). The rs2551188 minor T-allele has no specific association with ROS within the Caucasian population, although the SNP has been repeatedly associated with prostate cancer in Caucasian men (Kato *et al*., 2018). Within literature few rs10012 minor C-allele and ROS associations within the Caucasian population is documented with most associations being in women with endometrium cancer and men with prostate cancer (Ashton *et al.,* 2010; Beuten *et al.,* 2008). Cavalieri and Rogan (2014) stated that deviations in the oestrogen metabolism pathway is observed in women diagnosed with breast cancer, thyroid cancer and endometrium cancer and in men with prostate cancer and non-Hodgkins lymphoma. Thus, genetic variations that cause changes in oestrogen metabolites in one of these above-mentioned cancers could potentially cause the other cancers in other individuals corresponding to gender specificity.

The heterozygous genotype (CT/Mm) of rs2551188 is well known to increase *CYP1B1* enzyme expression in Caucasian population (Kato *et al*., 2018). For rs10012 it is well known that the homozygous minor C-allele genotype caused an amino acid substitution from Arg to Gly that substantially increases the *CYP1B1* catalytic activity (Ashton *et al.,* 2010). The increase in *CYP1B1* enzyme activity, in theory, leads to an increase of the formation of CE and consequently, an increase in ROS as claimed by Cavalieri and Rogan (2014) which does not correspond to our association analysis. Although the findings of Ashton *et al* (2010) indicate that a single variant within a haplotype with various SNPs, including rs10012, increases *CYP1B1* activity as well as decreases endometrium cancer risks which does correspond with our lowered ROS levels. Usually, the inhibition or lowered *CYP1B1* catalytic activity could decrease ROS production as Li *et al.* (2017) stated, the other possibility is that other 2 or 16-hydroxylated pathways could also be upregulated, and this lowered the substrates available for an increased catalytic activity of *CYP1B1*. The cooperation of SNPs within haplotypes could also lead to varying results that would differ to studies that investigates SNPs separately.

The minor allele for both SNPs rs2470890 (A > G) and rs2472304 (T > C) of *CYP1A2* demonstrated to be dominant in effect and these SNPs form a haplotype according to Figure 4.3 (a). The minor G-allele of rs2470890 and minor C-allele of rs2472304 associated with

increased levels of oestriol. There is no literature specifically stating an association between the SNPs rs2470890 or rs2472304 and oestriol in the Caucasian population. For some SNPs within the Caucasian population the heterozygous genotype (Mm) associated with a slight increase in oestriol levels. However, Vilčková *et al* (2023) stated that the rs2470890 homozygous minor G-allele genotype associated with decreased risk in prostate cancer in the Slovak population. This may be of relevance to this study since Balanovsky *et al*. (2008) stated that Russian genetics and allele distribution are very similar to Caucasian populations genetics. Furthermore, a Chinese population study by Bai *et* al (2017) stated that the minor G-allele of rs2470890 can co-dominantly or dominantly decrease the expression on the *CYP1A2* enzyme which corresponds to this association allele model. The decrease of *CYP1A2* activity could possibly increase alternative 16-hydroxylation of oestradiol and increase the oestriol levels as corresponds with our study and is more thoroughly discussed after the rs2472304 discussion section (Eliassen *et al.,* 2012) (Fig 2.2).

The other SNP, rs2472304, forms a common haplotype, especially in the Caucasian population, called CYP1A2*1F that is known to critically increase CYP1A2 activity (Lurie *et al.,* 2005*)*. Fekete *et al*. (2022) explains that the practical effect of a single SNP on *CYP1A2* activity could be different than the combined haplotype effect. Even though the haplotype is known to increase *CYP1A2* activity, specifically rs2472304 minor allele is associated with a lower 2-hydroxyestrone/16-hydroxyestrone ratio, indicating decreased *CYP1A2* catalytic activity and an upregulated alternative 16-hydroxylation pathway in the Caucasian population (Fekete *et al.,* 2022).

Therefore, both SNPs appear to decrease the *CYP1A2* expression or catalytic activity. The presence of both these SNPs in a haplotype may lead to the *CYP1A2* enzymes to be oversaturated with subsequent upregulation of the secondly preferred, 16-hydroxylation pathway (Eliassen *et al.,* 2012) (Fig 2.2). This could increase the protective oestriol hormone levels that corresponds to our association study as well as slightly decrease the more oestrogenic-active oestradiol hormone levels.

Eleven associations in the Caucasian population were grouped together since six of the seven *COMT* SNPs formed a haplotype according to Figure 4.3 (a). All seven SNPs associated with a significant decrease in both or one of the following oestriol-conjugates: oestriol-3-sulphate and oestriol-16-glucuronide. All the SNPs associated with heterozygous genotype model (dominant minor allele) rs4818 (C > G), rs4646312 (T > C), rs6269 (A > G), rs2239393 (A > G), rs4633 (T > C), rs4680 (A > G) and rs165599 (A > G). The first four SNPs, rs4818, rs4646312, rs6269 and rs2239393 associated in particular to both the decreased levels of the oestriol conjugates (oestriol-3-sulphate and oestriol-16-glucuronide). These four SNPs have no previously reported association with neither oestriol-3-sulphate, nor oestriol-16-glucuronide

as well as no associations reported for even the oestriol hormone within the Caucasian populations. These SNPs effects on *COMT* expression or activity are documented in a few articles mostly of *COMT* dependent pain studies. The soluble form of *COMT* (S-*COMT)* that degrade neurotransmitters relating to pain perception is also expressed within the liver (Yager, 2012). Kambur and Männistö (2010) also confirm that a SNP that associates with increased pain sensation is due to high amounts of neurotransmitters that can relay pain stimulation due to lower expression or catalytic activity of S-*COMT*.

SNP rs4818 minor G-allele has been associated with increased pain perception and is part of the high pain sensitivity (HPS) haplotype (Meloto *et al*., 2015). SNP rs4646312 minor C-allele associated with increased pain sensitivity, which might be due to a decreased expression or activity of *COMT* (Kambur & Männistö, 2010). rs6269 and rs2239393 are also part of the HPS haplotype with a hypothesis in the decrease of *COMT* expression (Tammimäki & Männistö, 2012).

These four SNPs are documented by their respective references to be synonymous, intron or 5' prime UTR located variants, which indicates differences in the secondary structure of mRNA of the *COMT* enzyme influencing the rate of enzyme expression (Kambur & Männistö, 2010). All these SNPs increase pain sensitivity and it is hypothesized that there is a decrease in expressed *COMT* enzymes (Diatchenko *et al*, 2005). A decrease in *COMT* enzymes could decrease methylated metabolites that interacted in a negative feedback loop with *CYP1A2*, consequently increasing 2-hydroxylation catalytic activities (Sak, 2017). The increase of 2-hydroxylation decreases the 16-hydroxylation and therefore the 16-hydroxylated metabolite conjugate as stated by Gorbach (1984) which corresponds to our haplotype and associations findings.

The following two associations of SNPs, rs4680 and rs4633, with decreased levels of oestriol-3-sulphate are still in haplotype with the four previously named SNPs but did not associate with decreased levels of oestriol-16-glucuronide. For rs4633 similar to the previous SNPs had also not been associated to oestriol within the Caucasian population but has been associated with increased risk in endometrium cancer (Hirata *et al*., (2008). rs4633 (T > C) is part of the HPS as mentioned by Diatchenko *et al.* (2005) and is reported to increase pain in Caucasian American females which hypothesises that this SNP decreases *COMT* expression (Andersen & Skorpen, 2009). The other SNP, rs4680, have multiple oestriol and breast cancer associations within literature (Sak, 2017). This rs4680 (A > G) SNP contradicts all five other SNPs in the haplotype and is determined to functionally increase *COMT* activity with a high activity minor G-allele and within this context a Met158Val amino acid substitution (Hall *et al*., 2016). Andersen and Skorpen (2009) emphasises the higher *COMT* activity of the minor G-allele since the SNP was also associated with decreased pain perception.

Two possible net hypotheses can arise from this phenomenon, the first is that the net effect of the haplotype is to decrease *COMT* expression regardless of rs4680 higher activity. The decrease of methoxy-oestrogens will negate the negative feedback loop *COMT* has with *CYP1A2* and this will upregulate the 2-hydroxylation pathway and indirectly cause the decrease of the oestriol-3-sulphate that would produce oestriol conjugates. The second net hypothesis is that the increased *COMT* activity caused by SNP rs4680 has a more prevalent effect than all the other SNPs within the haplotype. Consequently, there would be an increase in 2 and 4-methoxy-oestrogens along with up-regulation of the 2-hydroxylation pathways (except for *CYP1A2* enzyme) that could decrease the secondary 16-hydroxylaton pathway and therefore the oestriol-3-sulphates. Both hypotheses might explain the decrease in conjugates as a result of the genetic variation in comparison to a wild type phenomenon.

Six of seven SNPs of *COMT* (rs4818, rs4646312, rs6269, rs2239393, rs4633 and rs4680) formed a haplotype. Of the 25 Caucasian Control study group samples, 11 of these samples were minor allele carriers (either heterozygous or homozygous for the minor allele) of the first four SNPs (rs4818, rs4646312, rs6269 and rs2239393) and had decreased levels of both oestriol-3-sulphate and oestriol-16-glucuronide. With regards to the last two SNPs (s4633 and rs4680) 13 out of 25 samples were minor allele carriers (an additional two either heterozygous or homozygous for the minor allele) and presented decreased levels of oestriol-3-sulphate. Those additional two samples displayed only minor alleles for rs4633 (C-allele) and rs4680 (G-allele) indicating that even if only two minor alleles in the entire HPS haplotype is present, a phenotypical effect of decreased metabolite levels can still be seen (Diatchenko *et al*, 2005). Peterson *et al.* (2010) considers that a single minor allele in any SNP within specific haplotypes could be sufficient to cause a detectable metabolic change.

Lastly, SNP rs165599 which associated with decreased oestriol-16-glucuronide is not in any LD with any SNP and even with increased sample size would unlikely form a haplotype with the other *COMT* SNPs. rs165599 is very understudied even within more detailed pain perception studies (Vetterlein *et al*., 2023). The clearest determined effect and association was found in Govil *et al* (2020) were this rs165599 (A > G) associated with increased thermal pain perception. Increased pain suggests decreased *COMT* expression and less methoxy-oestrogens in the liver. The negative feedback loop for *CYP1A2* will decrease and the 2-hydroxylation pathway could be up-regulated, decreasing the oestriol-16-glucuronide levels which corresponds to our results.

5.1.2 African Control population

Within the African Control population group, five out of 29 associations were significant with three SNPs (rs2246709 in *CYP3A4*, and rs4680 and rs4633 in *COMT*) associating with three metabolites (2-hydroxyoestrone, 4-hydroxyoestrone and ROS) as shown in Table 4.8. None of the three associated SNPs formed haplotypes according to Figure 4.3 (b). Low haplotype frequency is to be suspected in the African population specifically with Xhosa descendants as stated by Masilela (2021) as well as a strict haplotype formation parameter of Haploview coupled with small sample sizes (Gabriel et al., 2002).

The first association was the dominant model of rs2246709 (A > G) which linked the minor G-allele with slightly increased ROS. Within literature rs2246709 has not been associated to ROS within the African population. This SNP's effect on enzyme activity is unknown and any effect has been accredited to another SNP within a known haplotype, namely rs35599367 G-allele (or alternatively complementary strand C-allele), that slightly increased *CYP3A4* expression (Wang *et al.,* 2011). Although within our African Control study group the rs3559367 had no minor allele carriers (either heterozygous or homozygous for the minor allele) whatsoever while rs2246709 had sufficient minor G-alleles which contrast the known African American haplotype as mentioned by Wang *et al.* (2011). This leads to a new hypothesis that an increase in *CYP3A4* activity in the absence of rs35599367 could instead be an effect of the rs2246709 SNP minor G-allele. An increased *CYP3A4* expression and subsequent increased hydroxylation would increase catechol oestrogens (CE) (Katzung, 2017). The increased CE leads to more quinone and semi-quinone CE that can partake in redox cycling to increase ROS formation, as stated by Cavalieri and Rogan (2014) that corresponds to our association analysis.

The following three associations in the African population is the minor A-allele as dominant within a heterozygous genotype of rs4680 (G > A) of *COMT* which associated with all three the metabolites with significantly lowered 2-hydroxyoestrone, 4-hydroxyoestrone and slightly lowered ROS. Rs4680 is a well-known SNP where the minor A-allele causes amino acid substitution of Valine to Methionine and results in a 4 to 5-fold lower *COMT* activity within the African American populations (Mondal, 2019). In correspondence the SNP has been associated with increased breast cancer risk and oestrogen-DNA adduct formation by Samavat and Kurzer (2015) and Mondal (2019) due to increased CE formation. Contradictory our results indicate a significant decrease in 2/4-hydroxyoestrone and slight ROS reduction that may be due to the 16-hydroxylation pathway upregulation to resolve the lower activity of *COMT.* If hydroxy-oestrogens decrease along with low *COMT* activity there could be an accumulation of oestrone and oestriol that will continue to have oestrogenic activity for longer

which could increase risk if oestrogen-related. The other possibility is that other surrounding factors cause a decrease in oestrogen that might be inherited with low activity rs4680 within African populations.

The 2-hydroxyoestrone is one of the five non-normally distributed metabolites within the African population group and was by chance one of the randomly selected for additional Mann-Whitney statistical analysis. The non-normal distributed 2-hydroxyoestrone according to the less influenced Mann-Whitney statistical analysis had a very slight increase in p-value, confirming the acceptable GLM statistical test tolerance to deviations from the normal distribution for the rest of this study.

The last association in the African population is the dominant, minor T-allele of rs4633 (C > T) of *COMT* which associated with slightly decreased ROS. This rs4633 SNP similarly to rs4680 associated with an oestrogen-related cancer, this case endometrium cancer and as well as having a minor T-allele that results in a lower *COMT* expression (Hirata *et al.,* 2008; Janacova, 2018). Our results contradict other studies association by having a slight decrease in ROS levels instead of increased levels that would be expected from such results. Within a pain focussed study where *COMT* degrades neurotransmitters, it was demonstrated by Sadhasivam *et al.* (2014) that rs4633 heterozygous genotype (CT) was associated with a decreased COMT activity in African American and European populations.

In this study it can be seen that the metabolites that are involved in pathways that may cause DNA damage are mostly associated with these selected SNP variants. The African population SNP effect on enzyme expression and catalytic activity as well as SNP-SNP interactions from the same or different genes still need to be further studied in order to understand how differences in metabolic pathways are induced.

5.1.3 Combined Controls population

Within the Combined Control population group, five out of 29 associations were significant with three SNPs (rs2246709 in *CYP3A4*, rs4818 and rs4646312 in *COMT*) associating with four metabolites (oestradiol-17-sulphate; testosterone; oestrone and oestriol-16-glucuronide) as shown in Table 4.8. All three associated SNPs formed no haplotypes according to Figure 4.3 (c). There is a low haplotype frequency due to the diversity from the African Controls within this study group (Masilela, 2021).

The first SNP, namely rs2246709 (A > G), has not yet been associated with increased oestradiol-17-sulphate or testosterone in the African population but has been in the Caucasian population (Elens *et al*., 2013). For both populations the minor G-allele is documented to

increase the *CYP3A4* activity compared to the other A-allele (Elens *et al.,* 2013; Wang *et al*, 2011). An increase in hydroxylation of *CYP3A4* increases the hydroxylated-oestradiol directly or after methylation to be sulfonated creating an increase in oestradiol-17-sulphate which corresponds to our results. The increased testosterone does not correspond to literature findings when the testosterone is used to test *CYP3A4* activity and should decrease hormone levels with increased enzyme activity (Elens *et al.,* 2013). The higher testosterone levels could be influenced by androgen biosynthesis enzymatic activities.

SNP rs4818 (C > G) minor G-allele associated with increases oestrone levels and decreased oestriol-16-glucuronide levels. R4818 has not yet been associated with any liver oestrogens within either population which corresponds with Hirata *et al.* (2008) stating that rs4818 is understudied compared to the other known *COMT* SNPs. Within the Caucasian population the rs4818 minor G-allele has been associated with breast cancer survivors as this increased *COMT* activity had a protective role against oestrogen (Udler *et al.,* 2009). In the study by Diatchenko *et al* (2005) the rs4818 associated with an increase in pain and therefore a decrease in *COMT* activity. The decrease in activity would explain the accumulation of oestrone and even the decrease in 16-hydroxylation conjugation since the 2-hydroxylation pathway could be up-regulated with the lack of methoxy-oestrogens for a negative feedback loop (Sak, 2017). Although the function of rs4818 minor allele on enzyme activity cannot be stated with certainty due to contradictory results demonstrated by Hirata *et al.* (2008) and Udler *et al.* (2009).

The last SNP is rs4646312 (T > C) minor C-allele associated with decreased levels of oestriol-16-glucuronide. This SNP also had no previous known association with oestriol or oestriol conjugates within either population groups. What is known is that rs4646312 minor C-allele (alternatively on the complementary strand known as the minor G-allele) is reported to associate with increased pain perception and thus indirectly decreased *COMT* activity (Kambur & Männistö, 2010). The reduction of methoxy-oestrogens could possibly up-regulate the 2-hydroxylation pathway and therefore decrease the 16-hydroxylation substrates which would correspond with our results.

5.1.4 Caucasian COC users

Within the COC group, four out of 29 associations were significant with two SNPs (rs1800440 and rs10012 both in *CYP1B1*) associating with four metabolites (16-ketoestradiol, an DNA-adduct, oestriol-16-glucuronide and 4-hydroxyoestrone) as shown in Table 4.8. Both rs1800440 (T > C) and rs10012 (G > C) minor alleles were shown to be dominant and it should be noted that these SNPs do not form a haplotype according to Figure 4.3 (a).

SNP rs1800440 minor C-allele associated with decreased levels of 16-Ketoestradiol, 4-OHE1-1N7G (DNA-oestrogen adduct) and oestriol-16-glucuronide. Rs1800440 is a polymorphism that causes an amino acid substitution from asparagine to serine. This substitution is known to cause an increased enzymatic activity even with decreased enzyme stability (Gajjar *et al.,* 2012). The minor C-allele has been associated with decreased hormone-related endometrium cancer within both Caucasian controls and contraceptive users (Ashton *et al*., 2010). An increased *CYP1B1* would increase 4-hydroxylation and thus decrease 16-hydroxylation metabolites such as 16-ketoestradiol and oestriol-16-glucuronide which corresponds to our findings. Although, one would expect the DNA-adduct, specifically 4-hydroxyoestrone-1-N-7-guanine, to also increase, this is not what we observed. A possible explanation could be the sufficient glutathione quinone conjugation that prevents the quinones cycling within a redox reaction and forming DNA-oestrogen adducts. Within another study done by Schults *et al* (2013) with Caucasian contraceptive users, the minor C-allele of rs1800440 also associated with a significant reduction in DNA-adducts indicating that metabolite levels can only be explained by genetics to a certain degree.

The last association of this study was between the rs10012 heterozygous genotype (GC) and increased 4-hydroxyoestrone. Ashton *et al* (2010) stated that rs10012 minor C-allele has an increasing catalytic effect on *CYP1B1.* As the *CYP1B1* activity increases the amount of 4-hydroxyoestrogens should also increase which corresponds with our findings.

5.2 Summary Chart

All the 29 associations from all the study groups are summarized within a single chart on the single landscape page 98 (97 in pdf document) (Figure 5.1). The SNPs are listed in a row at the top of the chart and are grouped per gene (*CYP1B1, CYP3A4, CYP1A2* and *COMT*) and the genes were arranged according to increasing chromosome number and the SNPs were arranged according to the position on the chromosome. The metabolites are listed in a column on the left-hand of the chart. Associations are displayed as a shape where the SNP and metabolite overlap within the middle of the chart. All associations with a p-value < 0.08 are displayed within the chart. The type of shape indicates which study group the association was measured in where the circle represents the Caucasian Controls, squares the African Controls, triangles the Combined Controls and lasty the rhombus as the Caucasian COC users. Another feature in this chart is the indication of increased metabolite levels by solid shapes and decreased metabolite levels by outlined shapes.

These discussions of the main results fully conclude Aim 2 of this study.

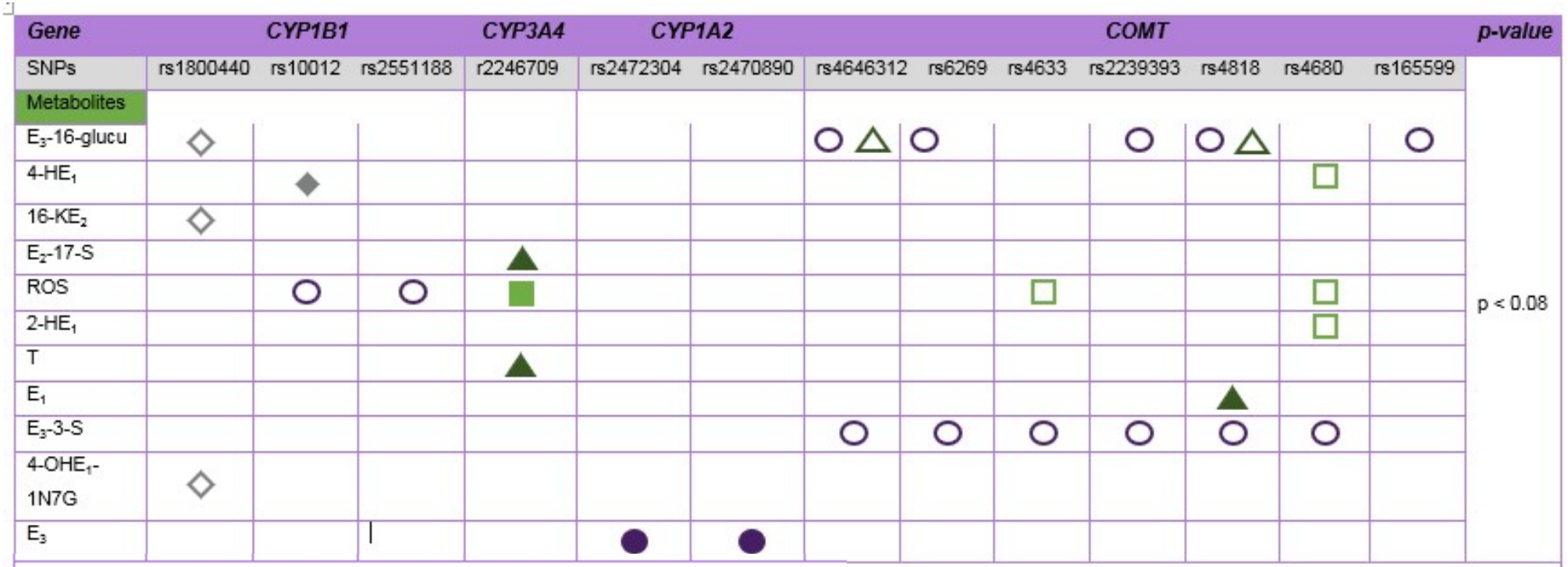


Figure 5.1 Summary chart of 29 associations

Legend

Caucasian	African	Combined Controls	COC users (Caucasian)	Increased metabolite levels	Decreased metabolites levels

CYP – Cytochrome; *COMT* – Catechol-o-methyltransferase; E_1 – Oestrone; E_2 – Oestradiol; E_3 – Oestriol; glucu – glucuronide; H – Hydroxy; K – Keto; S – Sulphate; ROS – Reactive oxygen species; T – Testosterone; 4-OHE1-1N7-G – oestrogen-DNA adduct; G – Guanidine

A summary of all 29 associations is displayed in this chart. The SNPs, grouped per gene, are listed in a top row and the metabolites are listed in the left-hand column and associations are displayed as a shape where the SNP and metabolites overlap within the middle of the chart. The shape type displays which study group the association was measured in as well as if the metabolite in the association increased (solid shape) or decreased (outlined shape).

5.3 SNP and Group Comparisons

Here, the Minor allele frequency (MAF) and minimum minor allele carriers (either heterozygous or homozygous for the minor allele) of the associating SNPs for all study groups are compared in Table 5.1. Furthermore, the measured association between two sets of different study groups (the Caucasian Controls and the African Controls; the Caucasian Controls and the Caucasian COC users) are displayed in boxplots in Figures 5.2 to 5.5.

5.3.1 SNP comparisons

In Table 5.1 only the MAFs of the SNPs that were significantly associated in this study, as shown in Table 4.8, are listed. Although there were 13 SNPs that associated with specific metabolites across the different population groups (section 5.1), the SNPs that associated were different per population group. All 13 SNPs are, however, listed here for all the population groups for the sake of interpretation and comparison. The data from these tables emphasizes the differences in SNP allele frequencies between the populations (Shifman *et al.,* 2003). The major and minor alleles that were applied to all the population groups for comparability with no bias to a specific study group, were determined by Ensembl (v109, human genome GRCh38) in which all population genetics are grouped together. However, as can be seen in Table 5.1, the minor allele identified by Ensembl “All” population group is not always the allele with the lowest frequency in a specific population group. The green highlighted MAF values in Table 5.1 indicate that the major and minor alleles in that population is actually opposite to that of the reference in Ensembl. The Caucasian and COC groups had five SNPs for which the chosen minor alleles had the highest frequency. It should be noted that MAF represents a percentage of the total alleles observed to be the minor allele and, on the other hand, “n” column represents the number of minor allele carriers (either heterozygous or homozygous for the minor allele) present [e.g., rs4646312 has a MAF of 0.40 (20 out of 50 alleles) and 16 minor allele carriers (either heterozygous or homozygous for the minor allele) out of 25 samples (four homozygous minor genotype and 12 heterozygous genotype that adds up to a total of 20 minor alleles).

The MAF of the SNPs of *CYP1B1* were similar between Caucasian and COC groups. The African and Combined controls had a higher MAF compared to the Caucasian groups except for rs1800440 that had no minor alleles within the African population and a lower MAF in the Combined group than in the Caucasian groups.

Table 5.1 Study Groups with Comparable SNP Frequencies

		Ensembl (1000 Genomes)		Alleles	Haploview (Study Groups)							
					Caucasian Controls (25 samples)		African Control (23 samples)		Combined Controls (48 samples)		COC (25 samples)	
Gene	SNP	African MAF	European MAF	Major : Minor	MAF (50 alleles)	n	MAF (46 alleles)	n	MAF (96 alleles)	n	MAF (50 alleles)	n
CYP1B1	rs1800440	<0.001	0.20	T : C	0.26	11	0.00	0	0.13	11	0.24	12
	rs10012	0.57	0.29	G : C	0.29	12	0.41	13	0.35	25	0.36	13
	rs2551188	0.52	0.29	C : T	0.28	12	0.39	13	0.34	25	0.36	13
CYP3A4	rs2246709	0.36	0.27	A : G	0.16	6	0.26	12	0.21	18	0.34	12
CYP1A2	rs2472304	0.03	0.60	G : A	0.64	22	0.11	5	0.39	27	0.60	23
	rs2470890	0.03	0.60	C : T	0.64	22	0.02	1	0.34	23	0.60	23
COMT	rs4646312	0.14	0.40	T : C	0.22	11	0.17	7	0.20	18	0.40	16
	rs6269	0.37	0.41	A : G	0.24	11	0.39	15	0.31	26	0.42	16
	rs4633	0.29	0.50	C : T	0.70	23	0.28	11	0.50	34	0.54	20
	rs2239393	0.41	0.41	A : G	0.24	11	0.44	17	0.33	28	0.42	16
	rs4818	0.17	0.40	C : G	0.22	11	0.26	10	0.24	21	0.40	16
	rs4680	0.28	0.5	G : A	0.70	23	0.22	10	0.47	33	0.54	20
	rs165599	0.25	0.69	G : A	0.74	25	0.22	7	0.49	32	0.72	23

Legend

COC - Combined oral contraceptive users (Caucasian); MAF - Minor Allele Frequency; n - The amount of samples that are Minor Allele carriers

CYP - Cytochrome P450; *COMT* – Catechol-O-methyltransferase

Purple highlighted samples were not significant

Green highlighted samples have opposite Major and Minor allele

Frequencies of all the SNPs with enough variance to make a significant association are listed within this table. There are four population groups listed above as well as Ensembl African and Caucasian documented MAF distributions, where the Ensembl “All” population group was used to determine the Major and Minor Alleles. The genes were arranged according to chromosome number and the SNPs were arranged according to the position on the chromosome. The SNPs MAF demonstrated the differences of the SNP distribution among the populations. The green highlighted MAF indicate that the Major and Minor allele are opposite of that demonstrated in the Ensembl “All” population group. All the purple highlighted regions in the "n" columns were SNPs with either insufficient sample sizes or non-significant associations.

With the single *CYP3A4* SNP, rs2246709, the African and Combined group MAFs were similar, although the Caucasian and COC users were quite different indicating how small sample size can skew frequencies that should be similar. For the *CYP1A1/1A2* SNPs the MAFs of the Caucasian and COC users were similar and the African MAF was significantly lower resulting in a lower Combined Control frequency. The *COMT* SNPs have various MAF differences such as the Caucasian and COC users being similar with rs165599 but COC having lower MAF with rs4633 and rs4680 and then higher frequencies with the remaining SNPs. The African Control group had lower frequencies with most SNPs except for rs6269 and rs2239393 which was higher than the Caucasian group.

5.3.2 Caucasian and African comparison

The associations found in the Caucasian and African populations groups were different and many factors could combine to result in such an outcome, such as SNP frequencies differences, lifestyles and diet. As seen in Figure 5.2 (a) the ROS metabolite has associated with five SNPs associations with two in the Caucasian population and the other three SNPs in the African population with no mutual over-lap in associated SNPs. African ROS levels were on average significantly higher than the Caucasian levels as mentioned in Table 4.9 in the Combined group. In the Caucasian study group rs10012 minor C-allele and rs2551188 minor T-allele associated with a decreased ROS level, while in the African population the ROS levels were not influenced by the SNPs, even though there were sufficient minor allele carriers (either heterozygous or homozygous for the minor allele). Although the minor G-allele of rs2246709 associated with increased ROS levels, it can be the major A-allele of rs2246709 along with two SNPs rs4680 minor A-allele and rs4633 minor T-allele associated with a decrease in ROS levels within the African study group.

In the Caucasian study group, rs4680 and rs4633 did not affect the ROS levels, even though they had sufficient minor allele carriers (either heterozygous or homozygous for the minor allele). The third SNP, rs2246709, had a significant HWE value in the Caucasian study group indicating that the minor allele frequency could be influence by one of many genetic factors and therefore cannot be unbiasedly used in association analysis.

Increased oestriol levels in Figure 5.2 (b) associated with the two SNPs, rs2470890 minor T-allele and rs2472304 minor A-allele, but only within the Caucasian population. The Caucasians rs2470890 major C-allele and rs2472304 major G-allele associated with higher oestriol levels than observed in the African study group. Furthermore, within the African population, these SNPs had an inverse minor/major allele frequency – the allele considered minor in the Caucasian population is actually the allele with the highest frequency within the

African population to such an extent that there was insufficient sample size containing the secondary allele to make an association within the African population.

In Figure 5.2 (c-d) the lower levels of 2- and 4-hydroxyoestrone associated with rs4680 minor A-allele within the African population. In the Caucasian population the major/minor alleles are inverse of that in the African population as seen in Table 5.1. Although the alleles are opposite and there was enough minor allele frequencies, the SNPs had no effect on Caucasian hydroxyoestrone levels as seen in Figure 5.2 (c-d).

In Figure 5.3 (a) the four SNPs, namely rs4646312 minor C-allele, rs6269 minor G-allele, rs4633 minor G-allele and rs2239393 minor G-allele associated with lower oestriol-16-glucuronide levels within the Caucasian population. Another fifth SNP, rs165599 has inverse major/minor alleles as displayed in Table 5.1 and lowest frequency G-allele associated with lower levels of oestriol-16-glucuronide. In the African population the levels of this specific oestriol conjugate did not vary much between samples and three of the five SNPs (rs4646312, rs6269 and rs2239393) had insufficient alleles to be used in association analysis.

In Fig 5.3 (b) the four SNPs, namely rs4646312 minor C-allele, rs6269 minor G-allele, rs2239393 minor G-allele and rs4818 minor G-allele, associated with lower oestriol-3-sulphate levels in the Caucasian population. The other two SNPs, rs4680 and rs4633 have inverse major/minor alleles as displayed in Table 5.1 and thus the lower frequency rs4680 G-allele and rs4633 C-allele associated with lower oestriol-3-sulphate levels. In the African population, the two SNPs, rs6269 major A-allele and rs2239393 major A-allele tended to have a higher oestriol-3-sulphate levels than in the Caucasian population, although no significant associations were found. The main reason could be that four of the six SNPs, rs4646312, rs4633, rs4818 and rs4680 in the African population did not influence the oestriol-3-sulphate levels and the other two SNPs, rs6269 and rs2239393 had insufficient minor allele frequencies. Other genes may elucidate the changes in these metabolite levels.

5.3.3 Caucasian and COC user comparison

The Caucasian Controls and the Caucasian COC users have the same genetic background but due to small sample sizes, the MAFs of the different SNPs were not always the same for the two groups. For some SNPs the COC group had insufficient frequencies and SNPs that were included in the Caucasian group association analysis had to be excluded for the COC group. Figure 5.4 (a) shows that the association of two SNPs (rs10012 minor C-allele and rs2551188 minor T-allele) decreased ROS level in the Caucasian controls was not seen in the COC users (the ROS levels did not differ between genotypes). The COC ROS levels were on average much higher than the Caucasian controls ROS levels.

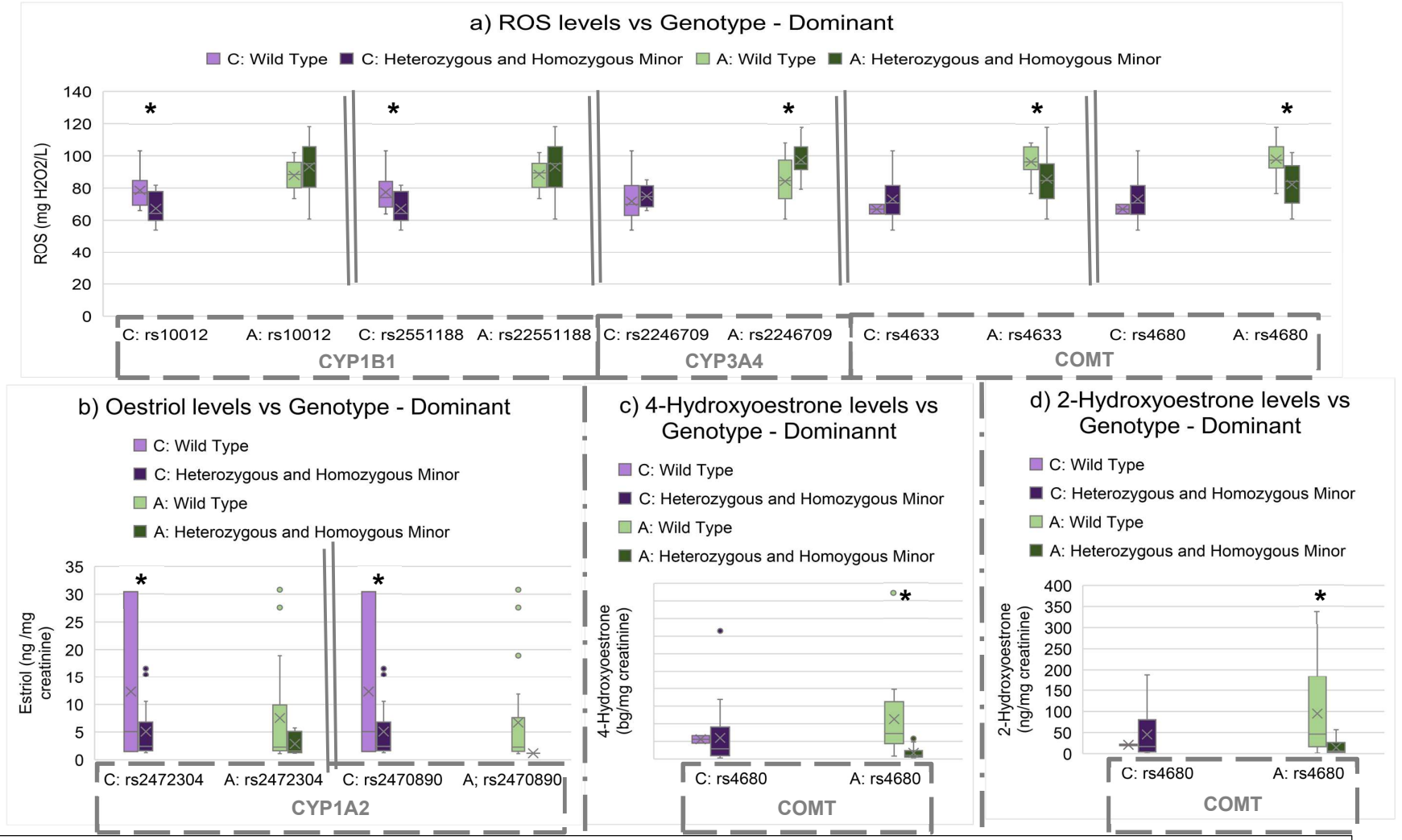


Figure 5.2 Multiple Boxplots comparing African and Caucasian SNP-metabolite associations

Legend

ROS – reactive oxygen species; C - Caucasian Controls; A – African Controls; *CYP* – Cytochrome; *COMT* – Catechol-o-methyltransferase

The boxplots are arranged per metabolite: (a) ROS, (b) oestriol, (c) 4-hydroxyoestrone and (d) 2-hydroxyoestrone as displayed in the title and y-axis. The boxplots compare significant SNP-metabolite associations between the Caucasian Controls (purple) and African Control (green) study groups in that order, per SNP. The SNPs on the x-axis are displayed per gene and is arranged according to the dominant model (major and minor allele determined by Table 5.1). In these boxplots some of the extreme outliers (when not in logged) were not included in order to display a graph with a visually noticeable change in metabolite.

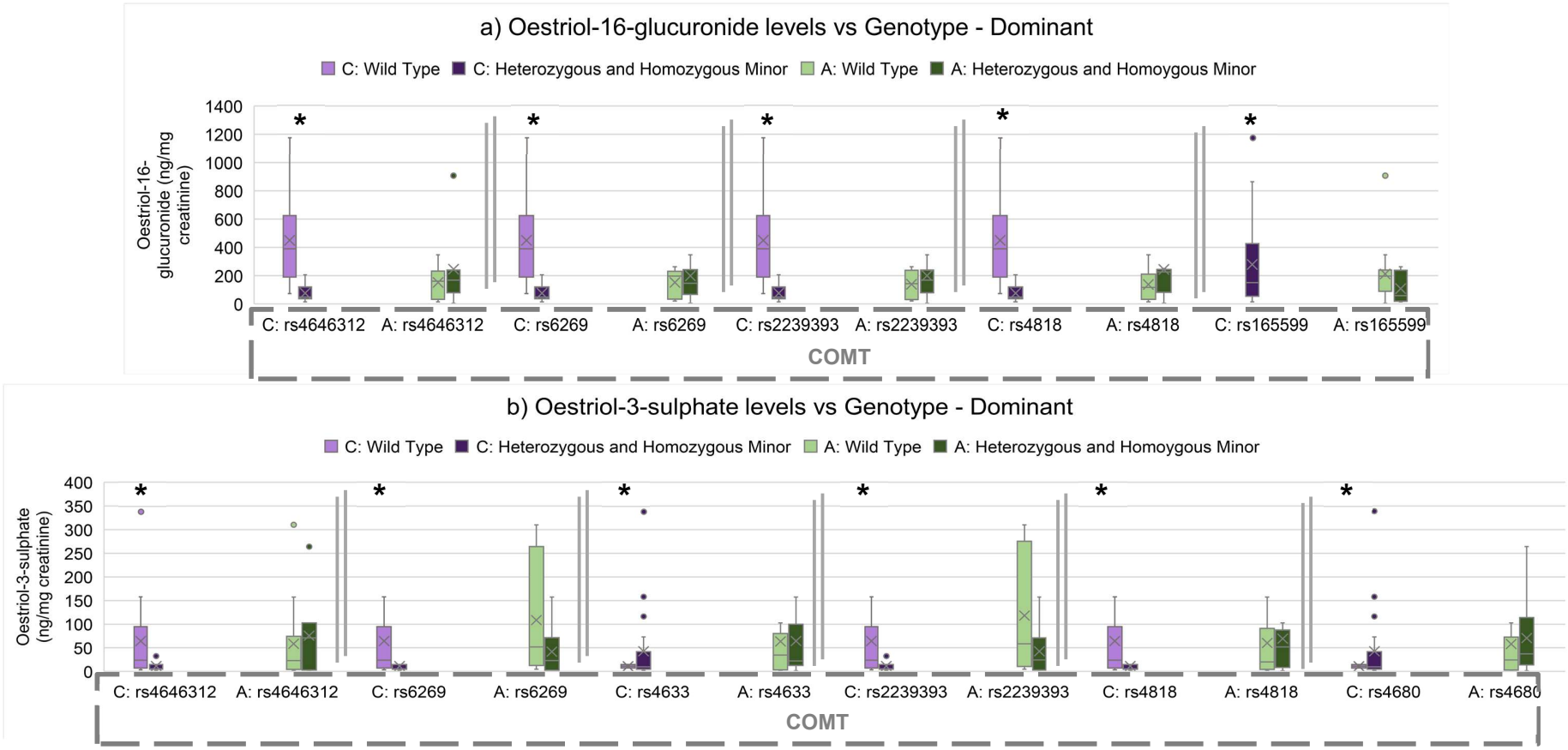


Figure 5.3 Multiple Boxplots comparing African and Caucasian SNPs-metabolite associations

Legend

C - Caucasian Controls; A – African Controls; *COMT* – Catechol-o-methyltransferase

The boxplots are arranged per metabolite: (a) oestriol-16-glucuronide and (b) oestriol-3-sulphate displayed in the title and y-axis. The boxplots compare significant SNP-metabolite associations between the Caucasian Controls (purple) and African Control (green) study groups in that order, per SNP. The SNPs on the x-axis are displayed per gene and is arranged according to the dominant model (major and minor allele determined by Table 5.1). In these boxplots some of the extreme outliers (when not in logged) were not included in order to display a graph with a visually noticeable change in metabolite.

In Figure 5.4 (b) an increase in 4-hydroxyoestrone levels associated with the minor C-allele of rs10012 for COC users and not the Caucasian control group.

Further in Figure 5.4 (c) the two *CYP1A2* SNPs have inverse major/minor allele as displayed in Table 5.1 and thus the less frequent rs2470890 minor C-allele and rs2472304 minor G-allele associated with increased levels of oestriol in the Caucasian control group. The COC study group did not have high enough minor allele frequencies in these SNPs to have been included in this association analysis. In Figure 5.4 (d) the 16-Ketoestradiol decreased in association with rs1800440 minor C-allele in the COC population while it seemed not to be influenced by genotype in the Caucasian control study group. In Figure 5.4 (e) the minor C-allele of the rs1800440 associated with lower levels of the DNA-adduct 4-hydroxyoestrone-1-N-7-Guanine in the COC study group, although again this phenomenon was not seen the Caucasian population.

Figure 5.5 (a) (as mentioned in 5.3.2 as well) shows that the four *COMT* SNPs, namely rs4646312 minor C-allele, rs6269 minor G-allele, rs4633 minor G-allele and rs2239393 minor G-allele associated with lower oestriol-16-glucuronide levels within the Caucasian population while these associations were not found in the COC study group. Another fifth SNP, rs165599 has inverse major/minor alleles as displayed in Table 5.1 and lowest frequency G-allele associated with lower levels of oestriol-16-glucuronide and this association was also not found in the COC study group.

However, the minor C-allele of rs1800440 did associate with decreased oestriol-16-glucuronide in the COC group, while it did not in the control group. Of the five SNPs that associated in the Caucasian control group, four of these SNPs (rs4646312, rs6269, rs4818 and rs2239393) has insufficient minor allele carriers (either heterozygous or homozygous for the minor allele) in the COC group for association analysis and the one SNP was tested (rs65599) did not show association with changing metabolite levels.

In Figure 5.5 (b) (as mentioned in 5.3.2 as well) the four SNPs, namely rs4646312 minor C-allele, rs6269 minor G-allele, rs2239393 minor G-allele and rs4818 minor G-allele, associated with lower oestriol-3-sulphate levels in the Caucasian control population. The other two SNPs, rs4680 and rs4633 have inverse major/minor alleles as displayed in Table 5.1 and thus the lower frequency rs4680 G-allele and rs4633 C-allele associated with lower oestriol-3-sulphate levels in the Caucasian control group. All these SNPs had insufficient minor allele frequencies in the COC group.

The Caucasian COC user study groups had on average higher ROS levels in comparison to the Caucasian controls since the increase of ethinyloestradiol (EE) from COC consumption could deplete antioxidant defence and increase the formation of ROS (Li *et al.,* 2023). COC

also had on average lower oestriol hormone, oestriol-conjugates and 2/4-hydroxyoestrones than the Caucasian control group. Drospirenone (DRSP), one component of COC, prevents pregnancy by triggering a negative feedback-loop that activates subsequent pathways that lead lowered biosynthesis oestrogens such as oestrone and oestriol (Cooper *et al.,* 2022). Fewer oestrogen metabolites synthesized could explain the decrease of oestriol and 2/4-hydroxyoestrones that is observed in COC users.

An interesting difference between Caucasian controls and COC users is the association between SNP rs1800440 and oestrogen metabolites from the 16-hydroxylation pathway(i.e., 16-Ketoestradiol and oestriol-16-glucuronide) that was not found in the control group. Li *et al*. (2023) states that the glucuronide and sulphate conjugates are primary part in EE metabolism and that there are 10-fold more EE-conjugates than EE metabolites indicating the lack of available glucuronide and sulphate for oestriol conjugation formation as well as other 16-hydroxylated metabolites. The rs1800440 SNP might still have the same effect in Caucasian controls although it can only be noticeably measured when the metabolites are placed under metabolism competition such as additional COC would provide.

Comparing African control and Caucasian controls, as well as Caucasian controls and Caucasian COC concludes objective 11 along with Aim 3.

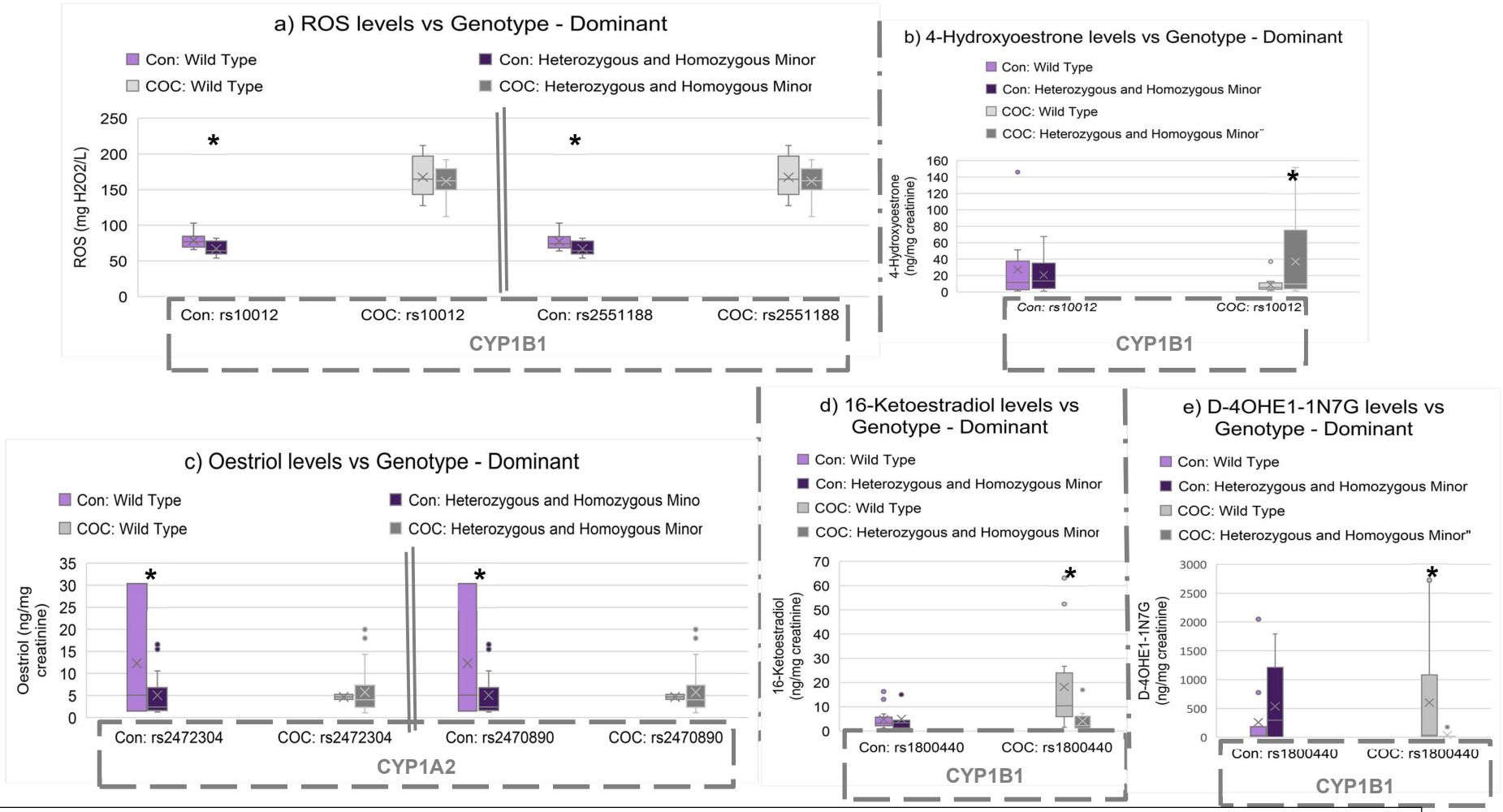


Figure 5.4 Multiple Boxplots comparing Caucasian Controls and Caucasian COC users SNPs-metabolite associations

Legend

ROS – reactive oxygen species; Con - Caucasian Controls; COC –Combined oral contraceptive users (Caucasian) ; *CYP* – Cytochrome

The boxplots are arranged per metabolite: (a) ROS, (b) 4-Hydroxyoestrone, (c) Oestriol, (d) 16-Ketoestradiol and (e) D-4OHE1-1N7G displayed in the title and y-axis. The boxplots compare significant SNP-metabolite associations between the Caucasian Controls (purple) and Caucasian COC users (grey) study groups in that order, per SNP. The SNPs on the x-axis are displayed per gene and is arranged according to the dominant model (major and minor allele determined by Table 5.1). In these boxplots some of the extreme outliers (when not in logged) were not included in order to display a graph with a visually noticeable change in metabolite.

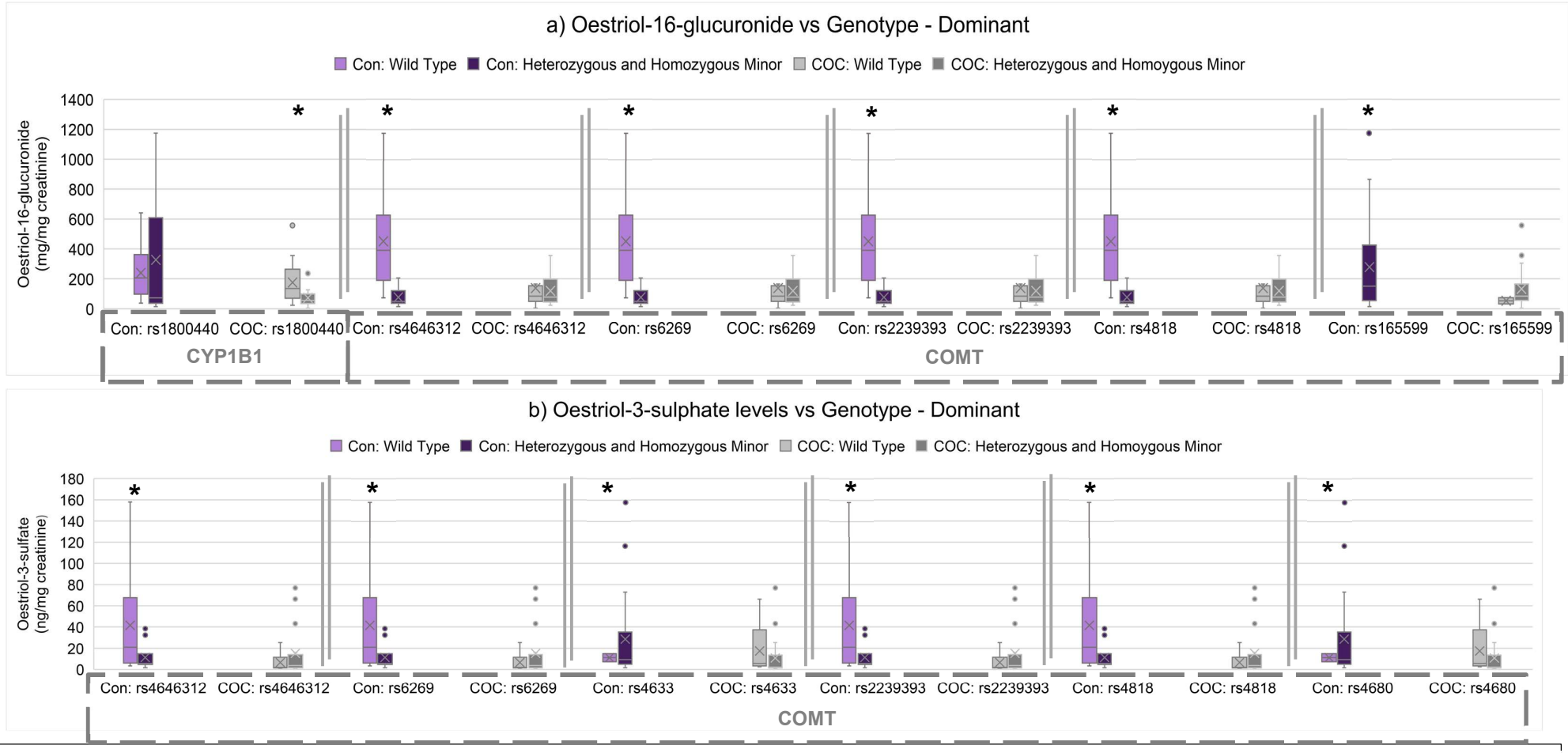


Figure 5.5 Multiple Boxplots comparing Caucasian Controls and Caucasian COC users SNPs-metabolite associations

Legend

ROS – reactive oxygen species; Con – Caucasian Controls; COC – Combined oral contraceptive users (Caucasian) ; *CYP* – Cytochrome; *COMT* – Catechol-o-methyltransferase

The boxplots are arranged per metabolite: (a) Oestrio-16-glucuronide and (b) Oestriol-3-sulphate in the title and y-axis. The boxplots compare significant SNP-metabolite associations between the Caucasian Controls (purple) and Caucasian COC users (grey) study groups in that order, per SNP. The SNPs on the x-axis are displayed per gene and is arranged according to the dominant model (major and minor allele determined by Table 5.1). In these boxplots some of the extreme outliers (when not in logged) were not included in order to display a graph with a visually noticeable change in metabolite.

CHAPTER 6 - CONCLUSION

It has been observed how a variation in genetic material did associate directly with changes in oestrogen metabolism levels within healthy, pre-menopausal South African women. Most of the SNPs listed in this study have documented known changes in activity or expression of an enzyme, although not always fully curated by enzymatic studies. SNP-metabolite associations were measured in four different study groups that included Caucasian and African ethnicities as well as the Caucasian study group with the included lifestyle choice of COC usage. Single SNP-metabolites associations as one of the predictors of metabolic changes with various external factor controls was essentially demonstrated in this study.

The list of metabolites obtained from the NWU eBOSS research group was successfully log transformed to display a slight, yet acceptable, deviation in normal curve distribution and was therefore used for association analyses. The genetic material was effectively isolated from eBOSS serum samples and genotyped at CPGR. A few SNPs that met the minimum requirements for downstream associating analyses were carefully selected from the numerous SNPs available in Haploview. The preliminary association analysis in PLINK realistically highlighted possible associations from multiple possibilities that was further studied in SPSS. The detailed association analyses results in SNP-metabolite associations proven to be significant was summarized in a reader-friendly chart. The complicated acquired associations were displayed in multiple boxplots in a comparative manner between either population groups or between Caucasians controls and Caucasian combined oral contraceptive (COC) users.

Even though most of the SNPs minor alleles have been associated with a wide metabolic-affecting disease, most of these SNP-metabolite specific associations are novel. The SNP-metabolite associations were done per study group, per metabolite and study groups had no overlap in association results. Within the Caucasian Control study group there were 15 SNP-metabolite associations which was the largest number of associations found of all the study groups and is briefly mentioned in this section. An increase in ROS levels associated with *CYP1B1* rs10012 minor C-allele and rs2551188 minor T-allele. An increase in oestriol (E_3) levels associated with *CYP1A2* rs2472304 minor C-allele and rs2470890 minor G-allele. A decrease in E_3-16-glucuronide associated with multiple *COMT* rs4818 minor G-allele, rs4646312 minor C-allele, rs6269 minor G-allele, rs2239393 minor G-allele and rs165599 minor G-allele. A decrease in E_3-3-sulphate associated with multiple *COMT* rs4818 minor G-allele, rs4646312 minor C-allele, rs6269 minor G-allele, rs2239393 minor G-allele, rs4633

minor C-allele and rs4680 minor G-allele. The Caucasian control population revealed a decrease in ROS and E_3-conjugates levels in the presence of SNP minor alleles as well as an increase in E_3 levels.

Within the African control study group there were five associations to be found in a population group that is understudied. An increase in ROS levels associated with *CYP3A4* rs2246709 minor G-allele. However, a decrease in ROS levels associated with the *COMT* rs4680 minor A-allele and rs4633 minor T-allele. The decrease of both 2/4 hydroxyoestrone (oestrone – E_1) levels associated with *COMT* rs4680 minor A-allele but not with rs4633. ROS can increase or decrease depending on which of the SNPs minor alleles are present. There is decrease in two and four hydroxylation of E_1, a product of phase I oestrogen metabolism in the presence of a phase II *COMT* minor allele SNP. It can be noted that the minor alleles for *COMT* rs4680 and rs4633 differ (are inverse).

Within the Combined controls study group with an increased sample size to a total of 48, five associations are found where ethnicity influence was considered. An increase in oestradiol (E_2)-17-sulphate and testosterone levels associated with *CYP3A4* rs2246709 minor G-allele. An increase in E_1 levels associated with *COMT* rs4818 minor G-allele. The decrease of E_3-16-glucuronide levels associated with *COMT* rs4818 minor G-allele and rs4646312 minor C-allele. The sufficient sample size highlights some non-population specific associations such as an increase in high oestrogen activity metabolites such as E_2 storage conjugate, E_1, Testosterone and a decrease in E_3-conjugate.

Within the Caucasian COC users study group, only four associations were found, the lowest number of all the study groups. A decrease in 16-KetoE_2, D-4OHE_1-1N7G (DNA-oestrogen adduct) and E_3-16-glucuronide levels all associated with *CYP1B1* rs1800440 minor C-allele. The other association was with the increase of 4-hydroxyE_1 and *CYP1B1* rs10012 minor C-allele. The decrease of E_3-16-glucuronide levels could be linked to the increase of 4-hydroxyE_1 levels which increases oestrogenic activity in the surrounding cells.

In the Caucasian control study group, the increase of *CYP1B1* enzyme activity/expression coupled with contradictory decrease of ROS levels is an example of the varying results haplotype interaction could cause. Also, if alternative 2/16-hydroxylation pathways are upregulated, the substrates available to *CYP1B1* will be reduced, leading to a protective effect of decreased ROS and increased E_3. In the African control group, the increased ROS levels along with decreased 2/4-hydroxyoestrone could indicate up-regulation of 4/16-hydroxylation pathway which results in more reactive catechol oestrogens (CE). In the Combined controls study group an increase of similar potency of oestrogenic activity metabolites (potential E_2 from E_2-storage conjugates, E_1 and testosterone) coupled with less oestrogenic effect,

decrease of E_3-conjugates levels are observed that could increase oestrogenic effect for longer periods. In the Caucasian COC user study group, the increase of 4-hydroxyE_1 levels could increase the formation of reactive CE which increases cancer risk.

Even though the BMI, age and menstrual phase were specifically selected to exclude these influences, specific few associations not listed in the sections above were removed after detailed association analyses in SPSS due to detected BMI and age influence. Also, the four metabolites ROS, FRAP, GSHt and Testosterone is determined to significantly differ between the Caucasian and African populations according to the detailed SPSS association analyses. It can be noted that the minor alleles for *COMT* rs4680 and rs4633 differ (are inverse). The difference in metabolite levels and SNP frequencies as well as the differences in SNP-metabolite associations highlight the differences between Caucasian and African population groups which impact pharmaceutical metabolism and effects. However, the Combined controls found SNP-metabolite associations that were applicable to both population groups.

The differences in metabolic levels of individuals within the same population group can be caused by a genetic variation represented by SNPs minor alleles. Other factors such as diet, environments and other genetic factors involved in protein expression could also interact to cause a varying effect on the metabolic levels. Additionally, the lifestyle factor of consuming COC is highlighted when comparing the Caucasian controls with the Caucasian COC users. The changes in metabolite levels from increased ethinylE_2 (EE) and drospirenone (DRSP) saturating enzymes revealed SNPs effect and therefore a measurable SNP-metabolite association. Also, the less used oestrogen metabolism pathways are upregulated in reaction to the changed metabolite levels, that could increase the formation of more reactive metabolites which increases cancer risk.

These samples were all acquired from healthy South African women, although throughout this section the change in metabolism as a result of genetic and/or lifestyle (COC consumption) factors could induce imbalance pathways. If these imbalanced pathways are frequently induced or become a “chronic” state, the increase of reactive metabolites levels could form DNA-oestrogen adducts and eventually induce a difficult to treat, cancerous diseases. In this study the initial step was taken to associate SNPs (that have in literature been associated with oestrogen-induced cancer patients) with changing metabolite levels in healthy women. When a SNP associates with a changing metabolite level, this possible genetic causality could lead to a better understanding in these differences in disease aetiology such as the most relevant breast cancer and therefore inspire slightly more accurate treatment or preferably prevention strategies.

For future studies I suggest confirming the ethnicity of the current samples by using ancestry informative markers (AIMS) (specific SNPs at certain frequencies) already genotyped in this study such as the 200+ African specific markers annotated on the Infinium GSA v2.0 beadchip files. Another suggestion is to study other genes, gene families and gene sub-families involved in oestrogen metabolism, biosynthesis or antioxidant mechanisms such as *SULT*, *GST*, *SOD*, other *CYP* (such as an African specific *CYP3A5)*, *HSD*, reductases, *NQO*. It is always beneficial to increase the sample size to have a minimum of 30 samples in any group at any given association in order to increase reliability of every measurement, although the study can continue with small sample sizes according to acceptable protocols. Also with increased sample sizes, haplotypes will more clearly form, which will enable more accurate observation of the interaction between SNPs in the same haplotype or SNP/gene interactions with other genes. If possible, to find a more specific writing style (maybe it already exists within literature) to refer to Caucasian and Black South Africans. An additional study could be to include other South African ethnicity controls and COC users as my last suggestion for future prospects.

REFERENCES

Ademuyiwa, F.O., Tao, Y., Luo, J., Weilbaecher, K. & Ma, C.X. 2017. Differences in the mutational landscape of triple-negative breast cancer in African Americans and Caucasians. *Breast cancer research and treatment*, 161: 491-499.

Adler, A.J., Wiley, G.B. & Gaffney, P.M. 2013. Infinium assay for large-scale SNP genotyping applications. *Journal of visualized experiments: JoVE,* (81).

Aklillu, E., Oscarson, M., Hidestrand, M., Leidvik, B., Otter, C. & Ingelman-Sundberg, M. 2002. Functional analysis of six different polymorphic CYP1B1 enzyme variants found in an Ethiopian population. *Molecular pharmacology*, 61(3): 586-594.

Alhusain, L. & Hafez, A.M. 2018. Nonparametric approaches for population structure analysis. *Human genomics*, 12(1): 1-12.

Alsubait, A., Aldossary, W., Rashid, M., Algamdi, A. & Alrfaei, B.M. 2020. CYP1B1 gene: Implications in glaucoma and cancer. *Journal of cancer*, 11(16):4652.

Amend, K., Hicks, D. & Ambrosone, C.B. 2006. Breast cancer in African-American women: differences in tumor biology from European-American women. *Cancer research*, 66(17): 8327-8330.

Andersen, S. & Skorpen, F. 2009. Variation in the COMT gene: implications for pain perception and pain treatment. *Pharmacogenomics* 10(4): 669-684.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. & Zondervan, K.T. 2010. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9): 1564-1573.

Anzai, Y., Heger-Mahn, D., Schellschmidt, I. & Marr, J. 2012. Suppression of ovarian activity with a low-dose 21/7-day regimen oral contraceptive containing ethinylestradiol 20 mcg/drospirenone 3 mg in Japanese and Caucasian women. *Contraception*, 86(1): 28-34.

Arslan, E., Tezcan, E., Camci, H. & Avci, M.K. 2021. Effect of DNA Concentration on Band Intensity and Resolution in Agarose Gel Electrophoresis. *Van Sağlık Bilimleri Dergisi*, 14(3): 326-333.

Aschard, H., Guillemot, V., Vilhjalmsson, B., Patel, C.J., Skurnik, D., Ye, C.J., Wolpin, B., Kraft, P. & Zaitlen, N. 2017. Covariate selection for association screening in multiphenotype genetic studies. *Nature genetics*, 49(12): 1789-1795.

Ashton, K.A. 2009. *Genetic Variation and Risk of Endometrial Cancer*. Newcastle: Medical Genetics University of Newcastle.

Ashton, K.A., Proietto, A., Otton, G., Symonds, I., McEvoy, M., Attia, J., Gilbert, M., Hamann, U. & Scott, R.J. 2010. Polymorphisms in genes of the steroid hormone biosynthesis and metabolism pathways and endometrial cancer risk. *Cancer Epidemiology*, 34(3): 328-337.

Badawi, A.F., Cavalieri, E.L. & Rogan, E.G. 2001. Role of human cytochrome P450 1A1, 1A2, 1B1, and 3A4 in the 2-, 4-, and 16 [alpha]-hydroxylation of 17 [beta]-estradiol. *Metabolism-Clinical and Experimental*, 50(9): 1001-1003.

Bai, X., Xie, J., Sun, S., Zhang, X., Jiang, Y. & Pang, D. 2017. The associations of genetic polymorphisms in CYP1A2 and CYP3A4 with clinical outcomes of breast cancer patients in northern China. *Oncotarget*, *8*(24): 38367.

Balanovsky, O., Rootsi, S., Pshenichnov, A., Kivisild, T., Churnosov, M., Evseeva, I., Pocheshkhova, E., Boldyreva, M., Yankovsky, N., Balanovska, E. & Villems, R. 2008. Two sources of the Russian patrilineal heritage in their Eurasian context. *The American Journal of Human Genetics*, 82(1): 236-250.

Barrett, J.C., Fry, B., Maller, J.D.M.J. & Daly, M.J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2): 263-265.

Bassoy, E.Y., Walch, M. & Martinvalet, D. 2021. Reactive oxygen species: Do they play a role in adaptive immunity?. *Frontiers in Immunology*, 12:55856.

Beuten, J., Gelfond, J.A., Byrne, J.J., Balic, I., Crandall, A.C., Johnson-Pais, T.L., Thompson, I.M., Price, D.K. & Leach, R.J. 2008. CYP1B1 variants are associated with prostate cancer in non-Hispanic and Hispanic Caucasians. *Carcinogenesis,* 29(9): 1751-1757.

Boesenberg-Smith, K.A., Pessarakli, M.M. & Wolk, D.M. 2012. Assessment of DNA yield and purity: an overlooked detail of PCR troubleshooting. *Clinical Microbiology Newsletter*, 34(1): 1-6.

Bolt, H.M. 1979. Metabolism of estrogens—natural and synthetic. *Pharmacology & therapeutics*, 4(1): 155-181.

Brody, T., 2016. Clinical Trials: Study Design, Endpoints and Biomarkers. *Drug Safety, and FDA and ICH Guidelines*, 847:304-305.

Brookes, A.J. 1999. The essence of SNPs. *Gene*, 234(2): 177-186.

Bu, Z.B., Ye, M., Cheng, Y. & Wu, W.Z. 2014. Four polymorphisms in the cytochrome P450 1A2 (CYP1A2) gene and lung cancer risk: a meta-analysis. *Asian Pacific journal of cancer prevention*, 15(14): 5673-5679.

Burton, L., Scaife, P., Paine, S.W., Mellor, H.R., Abernethy, L., Littlewood, P. & Rauch, C. 2020. Hydrostatic pressure regulates CYP1A2 expression in human hepatocytes via a mechanosensitive aryl hydrocarbon receptor-dependent pathway. *American Journal of Physiology-Cell Physiology*, 318(5): C889-C902.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S. & Munafò, M.R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience,* 14*(5):* 365*-376.*

Byrne, B.M. & Van de Vijver, F.J. 2010. Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International journal of testing*, *10*(2): 107-132*.*

Byrne, B.M. 2016. *Structural equation modeling with AMOS: Basic concepts, applications,* and programming. 1st ed. NY: Routledge

Calderón Giraldo, J. 2022. Analysis of estrogen profiles including methoxyestrogen glucuronides: method validation and applicability to human plasma and breast tissue. Würzburg: Julius-Maximilians-Universität Würzburg (Thesis – PhD).

Calitz, C., Hamman, J.H., Fey, S.J., Wrzesinski, K. & Gouws, C. 2018. Recent advances in three-dimensional cell culturing to assess liver function and dysfunction: from a drug biotransformation and toxicity perspective. *Toxicology mechanisms and methods*, 28(5): 369-385.

CANSA (The Cancer Association of South Africa). 2017. *The Big 5 Cancers affecting Women in South* Africa. https://cansa.org.za/infographic-the-big-5-cancers-affecting-women-in-sa/.Date of access: 25 Jul. 2020.

Carlsson, N., 2013. *Spectroscopic studies of biomolecules confined in self-assembled nanostructures*. Gothenburg: Chalmers University of Technology (Thesis – PhD).

Cauci, S., Buligan, C., Marangone, M. & Francescato, M.P. 2016. Oxidative stress in female athletes using combined oral contraceptives. *Sports medicine-open*, *2*: 1-9.

Cavalieri, E. & Rogan, E. 2014. The molecular etiology and prevention of estrogen-initiated cancers: Ockham's Razor: Pluralitas non est ponenda sine necessitate. Plurality should not be posited without necessity*. Molecular aspects of medicine*, 36: 1-55.

Cavalieri, E.L. & Rogan, E.G. 2016. Depurinating estrogen-DNA adducts, generators of cancer initiation: their minimization leads to cancer prevention. *Clinical and translational medicine*, *5*: 15.

Chainy, G.B. & Sahoo, D.K. 2020. Hormones and oxidative stress: an overview. *Free Radical Research*, *54*(1): 1-26.

Chakraborti, S. & Van de Wiel, M.A. 2008. A nonparametric control chart based on the Mann-Whitney statistic. In Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen. *IMS Colletions* (1): 156-173

Chen, J., Zhao, K.N. & Chen, C. 2014. The role of CYP3A4 in the biotransformation of bile acids and therapeutic implication for cholestasis. *Annals of translational medicine*, 2(1).

Cheng, K.C., Cahill, D.S., Kasai, H., Nishimura, S. & Loeb, L.A. 1992. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes GT and AC substitutions. *Journal of Biological Chemistry*, *267*(1): 166-172.

Cisneros, K.V., Agarwal, V. & James, M.O. 2019. Sulfonation and glucuronidation of hydroxylated bromodiphenyl ethers in human liver. *Chemosphere*, *226*: 132-139.

Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P. & Zondervan, K.T. 2011. Basic statistical analysis in genetic case-control studies. *Nature protocols*, *6*(2): 121-133.

CPGR (Centre for Proteomic & Genomic Research). 2022. Genomics. https://www.cpgr.org.za/services/genomics/ Date of access: 20 Jul. 2022.

Cooper, D.B., Patel, P., Mahdy, H. 2022. *Oral Contraceptive Pills.* Available from StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK430685/ Date of access: 17 Nov. 2023.

Curtis, K. & Youngquist, S.T. 2013. Part 21: categoric analysis: Pearson chi-square test. *Air medical journal*, *32*(4): 179-180.

Danielson, P.Á. 2002. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. *Current drug metabolism*, 3(6): 561-597.

Dawson, P.A., Elliott, A. & Bowling, F.G. 2015. Sulphate in pregnancy. *Nutrients*, *7*(3): 1594-1606.

De Leo, V., Musacchio, M.C., Cappelli, V., Piomboni, P. & Morgante, G. 2016. Hormonal contraceptives: pharmacology tailored to women's health. *Human reproduction update*, *22*(5): 634-646.

de Winter, J.C., Dodou, D.I.M.I.T.R.A. & Wieringa, P.A. 2009. Exploratory factor analysis with small sample sizes. *Multivariate behavioral research*, 44(2): 147-181.

Diatchenko, L., Slade, G.D., Nackley, A.G., Bhalang, K., Sigurdsson, A., Belfer, I., Goldman, D., Xu, K., Shabalina, S.A., Shagin, D. & Max, M.B. 2005. Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Human molecular genetics*, *14*(1): 35-143.

Drögemöller, B., Plummer, M., Korkie, L., Agenbag, G., Dunaiski, A., Niehaus, D., Koen, L., Gebhardt, S., Schneider, N., Olckers, A. & Wright, G. 2013. Characterization of the genetic variation present in CYP3A4 in three South African populations. *Frontiers in genetics*, *4*: 17.

Dubey, R.K. & Jackson, E.K. 2001. Estrogen-induced cardiorenal protection: potential cellular, biochemical, and molecular mechanisms. *American Journal of Physiology-Renal Physiology*, *280*(3): F365-F388.

Dumas, I. & Diorio, C. 2011. Estrogen pathway polymorphisms and mammographic density. *Anticancer research*, *31*(12): 4369-4386.

Eksteen, A. 2019. *Investigating the role of inflammation, progestins and steroid receptors in breast cancer*. Stellenbosch: Stellenbosch University) (Thesis – PhD).

Elens, L., Van Gelder, T., Hesselink, D.A., Haufroid, V. & Van Schaik, R.H. 2013. CYP3A4* 22: promising newly identified CYP3A4 variant allele for personalizing pharmacotherapy. *Pharmacogenomics*, *14*(1): 47-62.

Eliassen, A.H., Missmer, S.A., Tworoger, S.S., Spiegelman, D., Barbieri, R.L., Dowsett, M. & Hankinson, S.E. 2006. Endogenous steroid hormone concentrations and risk of breast cancer among premenopausal women. *Journal of the National Cancer Institute*, 98(19): 1406-1415.

Eliassen, A.H., Spiegelman, D., Xu, X., Keefer, L.K., Veenstra, T.D., Barbieri, R.L., Willett, W.C., Hankinson, S.E. & Ziegler, R.G. 2012. Urinary Estrogens and Estrogen Metabolites and Subsequent Risk of Breast Cancer among Premenopausal Women Estrogen Metabolites and Breast Cancer Risk. *Cancer research*, *72*(3): 696-706.

Faber, M.S., Jetter, A. & Fuhr, U. 2005. Assessment of CYP1A2 activity in clinical practice: why, how, and when?. *Basic & clinical pharmacology & toxicology*, *97*(3): 125-134.

Falah, N., Torday, J., Quinney, S.K. & Haas, D.M. 2015. Estriol review: Clinical applications and potential biomedical importance. *Clin Res Trials*, *1*(2): 29-33.

Fekete, F., Mangó, K., Minus, A., Tóth, K. & Monostory, K. 2022. CYP1A2 mRNA Expression Rather than Genetic Variants Indicate Hepatic CYP1A2 Activity. *Pharmaceutics*, *14*(3): 532.

Felty, Q., Xiong, W.C., Sun, D., Sarkar, S., Singh, K.P., Parkash, J. & Roy, D. 2005. Estrogen-induced mitochondrial reactive oxygen species as signal-transducing messengers. *Biochemistry*, *44*(18): 6900-6909.

Field, A. 2013. *Discovering statistics using IBM SPSS statistics*. 4th ed. LA: Sage.

Finco, A., Belcaro, G. & Cesarone, M.R. 2011. Assessment of the activity of an oral contraceptive on the levels of oxidative stress and changes in oxidative stress after co-treatment with two different types of physiological modulators with antioxidant action. *Contraception*, *84*(4): 418-422.

Fitzpatrick, D., Pirie, K., Reeves, G., Green, J. & Beral, V. 2023. Combined and progestagen-only hormonal contraceptives and breast cancer risk: A UK nested case–control study and meta-analysis. *Plos Medicine*, *20*(3): 1004188.

Flickinger, M., Jun, G., Abecasis, G.R., Boehnke, M. & Kang, H.M. 2015. Correcting for sample contamination in genotype calling of DNA sequence data. *The American Journal of Human Genetics*, 97(2): 284-290.

Frayling, T.M. 2014. Genome-wide association studies: the good, the bad and the ugly. *Clinical medicine*, *14*(4): 428.

Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. 2009. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, *10*(4): 241-251.

Fuller, K.N., McCoin, C.S., Stierwalt, H., Allen, J., Gandhi, S., Perry, C.G., Jambal, P., Shankar, K. & Thyfault, J.P. 2022. Oral combined contraceptives induce liver mitochondrial reactive oxygen species and whole-body metabolic adaptations in female mice. *The Journal of Physiology*, *600*(24): 5215-5245.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. & Liu-Cordero, S.N., 2002. The structure of haplotype blocks in the human genome. *Science*, 296(5576): 2225-2229.

Gajjar, K., Martin-Hirsch, P.L. and Martin, F.L. 2012. CYP1B1 and hormone-induced cancer. *Cancer letters*, *324*(1): 13-30.

Garson, G.D. 2012. *Testing statistical assumptions.* 1st ed. NC: Statistical Publishing Associates.

Gaudet, M.M., 2005. *Interactions of lifestyle factors, manganese superoxide dismutase, catechol-O-methyltransferase, and the risk of breast cancer*. Chapel Hill: The University of North Carolina at Chapel Hill (Thesis – PhD).

Gierisch, J.M., Coeytaux, R.R., Urrutia, R.P., Havrilesky, L.J., Moorman, P.G., Lowery, W.J., Dinan, M., McBroom, A.J., Hasselblad, V., Sanders, G.D. & Myers, E.R. 2013. Oral contraceptive use and risk of breast, cervical, colorectal, and endometrial cancers: a systematic review. *Cancer epidemiology, biomarkers & prevention*, *22*(11): 1931-1943.

Gorbach, S.L. 1984. Estrogens, breast cancer, and intestinal flora. *Reviews of Infectious Diseases*, 6(Supplement_1): S85-S90.

Govil, M., Mukhopadhyay, N., Holwerda, T., Sluka, K., Rakel, B. & Schutte, D.L. 2020. Effects of genotype on TENS effectiveness in controlling knee pain in persons with mild to moderate osteoarthritis. *European Journal of Pain*, 24(2): 398-412.

Guengerich, F.P. 2003. Cytochromes P450, drugs, and diseases. *Molecular interventions*, *3*(4): 194.

Hall, K.T., Jablonski, K.A., Chen, L., Harden, M., Tolkin, B.R., Kaptchuk, T.J., Bray, G.A., Ridker, P.M., Florez, J.C., Mukamal, K.J. & Chasman, D.I. 2016. Catechol-O-methyltransferase association with hemoglobin A1c. *Metabolism*, 65(7): 961-967.

Hertzog, M.A. 2008. Considerations in determining sample size for pilot studies. *Research in nursing & health*, 31(2): 180-191.

Hilborn, E., Stål, O. & Jansson, A. 2017. Estrogen and androgen-converting enzymes 17β-hydroxysteroid dehydrogenase and their involvement in cancer: with a special focus on 17β-hydroxysteroid dehydrogenase type 1, 2, and breast cancer. *Oncotarget*, 8(18): 30552.

Hirata, H., Hinoda, Y., Okayama, N., Suehiro, Y., Kawamoto, K., Kikuno, N., Rabban, J.T., Chen, L.M. & Dahiya, R. 2008. COMT polymorphisms affecting protein expression are risk factors for endometrial cancer. *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center*, *47*(10): 768-774.

Hobkirk, R. 1993. Steroid sulfation: current concepts. *Trends in Endocrinology & Metabolism*, *4*(2): 69-74.

Hodges, R.E. & Minich, D.M. 2015. Modulation of metabolic detoxification pathways using foods and food-derived components: a scientific review with clinical application. *Journal of nutrition and metabolism*, 2015(1).

Huo, D., Ikpatt, F., Khramtsov, A., Dangou, J.M., Nanda, R., Dignam, J., Zhang, B., Grushko, T., Zhang, C., Oluwasola, O. & Malaka, D. 2009. Population differences in breast cancer: survey in indigenous African women reveals over-representation of triple-negative breast cancer. *Journal of clinical oncology*, *27*(27): 4515.

Illumina, I. 2010. Infinium genotyping data analysis–a guide for analyzing infinium genotyping data using the genomestudio genotyping module.

Illumina, 2018. *Infinium™ Global Screening Array-24 v2.0, Technical notes.* https://countingchromosomes.com/blogfiles/2019-08-17/Illumina-GSA-data-sheet-370-2016-016.pdf Date of access: 16 Sep. 2022

Janacova, B.L. 2015. *Novel prometastatic proteins in breast cancer and their molecular role*. Brno: Masaryk University (Thesis – PhD).

Jiang, X., Waterland, M., Blackwell, L., Wu, Y., Jayasundera, K.P. & Partridge, A. 2009. Sensitive determination of estriol-16-glucuronide using surface plasmon resonance sensing. *Steroids*, *74*(10-11): 819-824.

Johansson, B.G. 1972. Agarose gel electrophoresis. Scandinavian *Journal of Clinical and Laboratory Investigation*, 29(sup124): 7-19.

Jumuddin, F.A., 2018. Antioxidant Properties of NQO2. Manchester: The University of Manchester (Thesis – PhD).

Kaabouch, N. and Schultz, R.R., 2007, January. A 2-D gel electrophoresis DNA image analysis algorithm with automatic thresholding. In *Visual Communications and Image Processing* 2007, 6508: 528-539

Kambur, O. & Männistö, P.T. 2010. Catechol-O-methyltransferase and pain. *International review of neurobiology*, *95*: 227-279.

Kato, I., Cichon, M., Yee, C.L., Land, S. & Korczak, J.F. 2009. African American-preponderant single nucleotide polymorphisms (SNPs) and risk of breast cancer. *Cancer epidemiology*, *33*(1): 24-30.

Kato, T., Hashimoto, Y., Wong, R.K., Mitsui, Y., Maekawa, S., Chang, I., Shahryari, V., Yamamura, S., Majid, S., Saini, S. & Tabatabai, Z.L. 2018. Influence of lifestyle choices on risks of CYP 1B1 polymorphisms for prostate cancer. *Journal of cellular and molecular medicine*, *22*(10): 4676-4687.

Katzung, B.G., 2017. *Basic and clinical pharmacology* 14th ed. NY: McGraw Hill

Key, T.J., Verkasalo, P.K. & Banks, E. 2001. Epidemiology of breast cancer. *The lancet oncology*, *2*(3): 133-140.

Key, T.J. 2011. Endogenous oestrogens and breast cancer risk in premenopausal and postmenopausal women. Steroids, 76(8): 812-815.

Kirpich, A., Ainsworth, E.A., Wedow, J.M., Newman, J.R., Michailidis, G. & McIntyre, L.M. 2018. Variable selection in omics data: A practical evaluation of small sample sizes. *PloS one*, *13*(6): e0197910.

Koetsier, G. & Cantor, E. 2019. A practical guide to analyzing nucleic acid concentration and purity with microvolume spectrophotometers, Technical Note. *New England Biolabs Inc*: 1-8.

Krauss, R.M. & Burkman Jr, R.T. 1992. The metabolic impact of oral contraceptives. *American journal of obstetrics and gynecology*, *167*(4): 1177-1184.

Lakowicz, J.R. 1999. Introduction to fluorescence. *Principles of fluorescence spectroscopy*: 1-23.

Lee, H.R., Kim, T.H. & Choi, K.C. 2012a. Functions and physiological roles of two types of estrogen receptors, ERα and ERβ, identified by estrogen receptor knockout mouse. *Laboratory animal research*, *28*(2): 71-76.

Lee, P.Y., Costumbrado, J., Hsu, C.Y. & Kim, Y.H. 2012b. Agarose gel electrophoresis for the separation of DNA fragments. JoVE (*Journal of Visualized Experiments*), (62): e3923.

Li, H.R. & Anderson, R.A. 2010. Recent advances in hormonal contraception. *F1000 medicine reports*, 2.

Li, J., Zhang, L., Zhou, H., Stoneking, M. & Tang, K. 2011. Global patterns of genetic diversity and signals of natural selection for human ADME genes. *Human molecular genetics*, 20(3): 528-540.

Li, J., Lao, X., Zhang, C., Tian, L., Lu, D. & Xu, S. 2014. Increased genetic diversity of ADME genes in African Americans compared with their putative ancestral source populations and implications for Pharmacogenomics. *BMC genetics*, 15(1): 1-15.

Li, F., Zhu, W. & Gonzalez, F.J. 2017. Potential role of CYP1B1 in the development and treatment of metabolic diseases. *Pharmacology & therapeutics*, *178*: 18-30.

Li, Z., Zhang, P., Yang, B., Liu, J., Xi, H., Zhang, D. & Yamaguchi, Y. 2021. High throughput DNA concentration determination system based on fluorescence technology. *Sensors and Actuators B: Chemical*, 328: 128904.

Li, L., Yang, X., Tran, D., Seo, S.K. & Lu, Y. 2023. Combined oral contraceptives as victims of drug interactions. *Drug Metabolism and Disposition*, *51*(6): 718-732.

Life Technologies. 2015. *Qubit® dsDNA BR Assay Kits*. https://www.thermofisher.com/order/catalog/product/Q32851 Date of access: 21 Sep. 2022

Liska, D., Lyon, M. & Jones, D.S. 2006. Detoxification and biotransformational imbalances. *Explore*, 2(2): 122-140.

Liu, Y. & Lu, L.Y. 2020. BRCA1 and homologous recombination: Implications from mouse embryonic development. *Cell & Bioscience*, 10(1): 49.

Long, J.R., Cai, Q., Shu, X.O., Cai, H., Gao, Y.T. & Zheng, W. 2007. Genetic polymorphisms in estrogen-metabolizing genes and breast cancer survival. *Pharmacogenetics and genomics*, 17(5): 331-338.

Lurie, G., Maskarinec, G., Kaaks, R., Stanczyk, F.Z. & Le Marchand, L. 2005. Association of genetic polymorphisms with serum estrogens measured multiple times during a 2-year period in premenopausal women. *Cancer Epidemiology Biomarkers & Prevention*, 14(6): 1521-1527.

Madhad, V.J. & Sentheil, K.P. 2014. The Rapid & Non-Enzymatic isolation of DNA from the Human peripheral whole blood suitable for Genotyping. *European Journal of Biotechnology and Bioscience*, 1(3): 01-16.

Männistö, P.T. & Kaakkola, S. 1999. Catechol-O-methyltransferase (COMT): biochemistry, molecular biology, pharmacology, and clinical efficacy of the new selective COMT inhibitors. *Pharmacological reviews*, 51(4): 593-628.

Masilela, C.M. 2021. *Investigation of socio-demographic, clinical and genetic factors associated with blood pressure and glycaemic control among indigenous South African adult patients.* CapeTown: University of the Western Cape (Thesis – PhD).

Mauvais-Jarvis, F., Clegg, D.J. & Hevener, A.L. 2013. The role of estrogens in control of energy balance and glucose homeostasis. *Endocrine reviews*, 34(3): 309-338.

Meloto, C.B., Segall, S.K., Smith, S., Parisien, M., Shabalina, S.A., Rizzatti-Barbosa, C.M., Gauthier, J., Tsao, D., Convertino, M., Piltonen, M.H. & Slade, G.D. 2015. COMT gene locus: new functional variants. *Pain*, 156(10): 2072.

Mills, J. 2023. A specialist in applicable molecular practices at Separations and Infinium Tech Support [email discussions]. Cape Town.

Mittler, R. 2017. ROS are good. *Trends in plant science*, 22(1): 11-19.

Mondal, B. 2019. *Modulation of Estrogen Metabolism and Prevention of Pathologies*. Nebraska: University of Nebraska Medical Center. (Thesis – PhD).

Moon, Y.J., Wang, X. & Morris, M.E. 2006. Dietary flavonoids: effects on xenobiotic and carcinogen metabolism. *Toxicology in vitro,* 20(2): 187-210.

Nakamura, J., Walker, V.E., Upton, P.B., Chiang, S.Y., Kow, Y.W. & Swenberg, J.A.. 1998. Highly sensitive apurinic/apyrimidinic site assay can detect spontaneous and chemically induced depurination under physiological conditions. *Cancer research*, 58(2): 222-225.

Nam, D., Kim, J., Kim, S.Y. and Kim, S., 2010. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic acids research*, 38(suppl_2): W749-W754.

Naz, F. 2014. *Biochemical and cytogenetic effects of oral contraceptives among women*. Aligarh: Aligarh Muslim University (Thesis – PhD).

Okoh, V., Deoraj, A. & Roy, D. 2011. Estrogen-induced reactive oxygen species-mediated signalings contribute to breast cancer. Biochimica et Biophysica Acta (BBA)-*Reviews on Cancer*, 1815(1): 115-133.

Oosting, J., Oosting, M.J. 2014. SNP biocViews Copy Number Variation, Genetic Variability and Preprocessing - Package 'beadarraySNP'.

Ozougwu, J.C. 2017. Physiology of the liver. *International Journal of Research in Pharmacy and Biosciences*, *4*(8): 13-24.

Peterson, N.B., Trentham-Dietz, A., Garcia-Closas, M., Newcomb, P.A., Titus-Ernstoff, L., Huang, Y., Chanock, S.J., Haines, J.L. & Egan, K.M. 2010. Association of COMT haplotypes and breast cancer risk in caucasian women. *Anticancer research*, 30(1): 217-220.

Pruthi, S., Yang, L., Sandhu, N.P., Ingle, J.N., Beseler, C.L., Suman, V.J., Cavalieri, E.L. & Rogan, E.G. 2012. Evaluation of serum estrogen-DNA adducts as potential biomarkers for breast cancer risk. *The Journal of steroid biochemistry and molecular biology*, 132(1-2): 73-79.

Qian, X., Xu, D., Liu, H., Lin, X., Yu, Y., Kang, J., Sheng, X., Xu, J., Zheng, S., Xu, D. & Qi, J. 2017. Genetic variants in 5p13. 2 and 7q21. 1 are associated with treatment for benign prostatic hyperplasia with the α-adrenergic receptor antagonist. *The Aging Male*, *20*(4): 250-256.

Quan, L., Hong, C.C., Zirpoli, G., Roberts, M.R., Khoury, T., Sucheston-Campbell, L.E., Bovbjerg, D.H., Jandorf, L., Pawlish, K., Ciupak, G. & Davis, W. 2014. Variants in estrogen-related genes and breast cancer risk in European and African American women. *Endocrine-related cancer*, 21(6): 853.

Raftogianis, R., Creveling, C., Weinshilboum, R. & Weisz, J. 2000. Chapter 6: Estrogen metabolism by conjugation. *JNCI Monographs*, 2000(27): 113-124.

Ray, P.D., Huang, B.W. & Tsuji, Y. 2012. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cellular signalling,* 24(5): 981-990.

Razali, N.M. & Wah, Y.B. 2011. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1): 21-33.

Reding, K.W., Chen, C., Lowe, K., Doody, D.R., Carlson, C.S., Chen, C.T., Houck, J., Weiss, L.K., Marchbanks, P.A., Bernstein, L. & Spirtas, R. 2012. Estrogen-related genes and their contribution to racial differences in breast cancer risk. *Cancer Causes & Control*, 23: 671-681.

Rendic, S. 2002. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug metabolism reviews*, 34(1-2): 83-448.

Rettberg, J.R., Yao, J. & Brinton, R.D. 2014. Estrogen: a master regulator of bioenergetic systems in the brain and body. *Frontiers in neuroendocrinology*, 35(1): 8-30.

Rodrigues, A.D. 2022. Drug Interactions Involving 17α-Ethinylestradiol: Considerations Beyond Cytochrome P450 3A Induction and Inhibition. *Clinical Pharmacology & Therapeutics*, 111(6): 1212-1221.

Roy, D. & Liehr, J.G. 1999. Estrogen, DNA damage and mutations. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 424(1-2): 107-115.

Ryckman, K. & Williams, S.M. 2008. Calculation and use of the Hardy-Weinberg model in association studies. *Current protocols in human genetics*, 57(1): 1-18.

Sadhasivam, S., Chidambaran, V., Olbrecht, V.A., Esslinger, H.R., Zhang, K., Zhang, X. & Martin, L.J. 2014. Genetics of pain perception, COMT and postoperative pain management in children. *Pharmacogenomics*, 15(3): 277-284.

Saini, A. 2021. Estrogen and Estrogen Receptor a Risk Factor in Breast Cancer. *International Journal of Biomedical and Advance Research*; 12(06): e5617.

Sak, K. 2017. The Val158Met polymorphism in COMT gene and cancer risk: role of endogenous and exogenous catechols. *Drug Metabolism Reviews*, *49*(1): 56-83.

Samavat, H. & Kurzer, M.S. 2015. Estrogen metabolism and breast cancer. *Cancer letters*, 356(2): 231-243.

Sansom, B. 2021. *Prevalence of genetic polymorphisms associated with anti-cancer drug efficacy and toxicity in the South African population*. Pretoria: University of Pretoria (Thesis – PhD)

Savage, K.I., Matchett, K.B., Barros, E.M., Cooper, K.M., Irwin, G.W., Gorski, J.J., Orr, K.S., Vohhodina, J., Kavanagh, J.N., Madden, A.F. & Powell, A. 2014. BRCA1 Deficiency Exacerbates Estrogen-Induced DNA Damage and Genomic Instability Estrogen-Induced Genomic Instability in BRCA1-Deficient Cells. Cancer research, 74(10): 2773-2784.

Schults, M.A., Chiu, R.K., Nagle, P.W., Wilms, L.C., Kleinjans, J.C., Van Schooten, F.J. & Godschalk, R.W. 2013. Genetic polymorphisms in catalase and CYP1B1 determine DNA adduct formation by benzo (a) pyrene ex vivo. *Mutagenesis*, 28(2): 181-185.

Sepkovic, D.W. & Bradlow, H.L. 2009. Estrogen hydroxylation—the good and the bad. *Annals of the New York Academy of Sciences*, *1155*(1): 57-67.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B. & Darvasi, A. 2003. Linkage disequilibrium patterns of the human genome across populations. *Human molecular genetics*, *12*(7): 771-776.

Singer, B.H., Ryff, C.D. & National Research Council. 2001. *Predisease Pathways. In New Horizons in Health: An Integrative Approach.* 1st ed. DC: National Academies Press (US).

Skibola, C.F., Bracci, P.M., Paynter, R.A., Forrest, M.S., Agana, L., Woodage, T., Guegler, K., Smith, M.T. & Holly, E.A. 2005. Polymorphisms and Haplotypes in the Cytochrome P 450 17A1, Prolactin, and Catechol-O-Methyltransferase Genes and Non-Hodgkin Lymphoma Risk. *Cancer Epidemiology Biomarkers & Prevention*, *14*(10): 2391-2401.

Smith, R.L. & Williams, R.T. 1970. *History of the discovery of the conjugation mechanisms. In Metabolic conjugation and metabolic hydrolysis*. 1st ed. Academic Press.

Sood, D., Johnson, N., Jain, P., Siskos, A.P., Bennett, M., Gilham, C., Busana, M.C., Peto, J., dos-Santos-Silva, I., Keun, H.C. & Fletcher, O. 2017. CYP3A7* 1C allele is associated with reduced levels of 2-hydroxylation pathway oestrogen metabolites. *British journal of cancer*, 116(3): 382-388.

Swart, M. & Dandara, C. 2014. Genetic variation in the 3′-UTR of CYP1A2, CYP2B6, CYP2D6, CYP3A4, NR1I2, and UGT2B7: potential effects on regulation by microRNA and pharmacogenomics relevance. *Frontiers in genetics*, 5: 167.

Taioli, E., Trachman, J., Chen, X., Toniolo, P. & Garte, S.J. 1995. A CYP1A1 restriction fragment length polymorphism is associated with breast cancer in African-American women. *Cancer Research*, 55(17): 3757-3758.

Tammimäki, A. & Männistö, P.T. 2012. Catechol-O-methyltransferase gene polymorphism and chronic human pain: a systematic review and meta-analysis. *Pharmacogenetics and genomics*, 22(9): 673-691.

ThermoFisher Scientific, 2021, *Invitrogen – Qubit Assay Quick Reference – MAN0017210,* https://manuals.plus/m/8456b47b134bd42f724c367dc35f0ada513ad1088d0c74b5fc7539b658ef7f37 Date of access: 14 Sep. 2022

Thomas, P.E., Klinger, R., Furlong, L.I., Hofmann-Apitius, M. & Friedrich, C.M. 2011. Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers. *BMC bioinformatics, 12: 1-18.*

Tian, X., Huo, X., Dong, P., Wu, B., Wang, X., Wang, C., Liu, K. & Ma, X. 2015. Sulfation of melatonin: enzymatic characterization, differences of organs, species and genders, and bioactivity variation. *Biochemical Pharmacology*, *94*(4): 282-296.

Trabert, B., Brinton, L.A., Anderson, G.L., Pfeiffer, R.M., Falk, R.T., Strickler, H.D., Sliesoraitis, S., Kuller, L.H., Gass, M.L., Fuhrman, B.J. & Xu, X. 2016. Circulating Estrogens and Postmenopausal Ovarian Cancer Risk in the Women's Health Initiative Observational Study Estrogen Metabolites and Ovarian Cancer Risk. *Cancer Epidemiology, Biomarkers & Prevention*, 25(4): 648-656.

Udler, M.S., Azzato, E.M., Healey, C.S., Ahmed, S., Pooley, K.A., Greenberg, D., Shah, M., Teschendorff, A.E., Caldas, C., Dunning, A.M. & Ostrander, E.A. 2009. Common germline polymorphisms in COMT, CYP19A1, ESR1, PGR, SULT1E1 and STS and survival after a diagnosis of breast cancer. *International journal of cancer*, 125(11): 2687-2696.

Van der Meer, Y. 2017. *Investigating progesterone and estrogen receptor crosstalk in breast cancer*. Stellenbosch: Stellenbosch University (Thesis – PhD).

Venter, G. 2021a. Ongoing inter- and intra-ethnic metabolite level differences in the biotransformation pathway conducted by the eBOSS research group [various in person discussions]. Potchefstroom.

Venter, G., van der Berg, C.L., van der Westhuizen, F.H. & Erasmus, E. 2021b. Health status is affected, and Phase I/II biotransformation activity altered in young women using oral contraceptives containing Drospirenone/Ethinyl Estradiol. *International Journal of Environmental Research and Public Health*, 18(20): 10607

Vetterlein, A., Monzel, M. & Reuter, M. 2023. Are catechol-O-methyltransferase gene polymorphisms genetic markers for pain sensitivity after all?. A review and meta-analysis. *Neuroscience & Biobehavioral Reviews*: 105112.

Vilčková, M., Škereňová, M., Dobrota, D., Kaplán, P., Jurečeková, J., Kliment, J., Híveš, M., Dušenka, R., Evin, D., Brožová, M.K. & Sivoňová, M.K. 2023. Polymorphisms in the gene encoding CYP1A2 influence prostate cancer risk and progression. *Oncology Letters*, 25(2): 1-9.

Wang, D., Guo, Y., Wrighton, S.A., Cooke, G.E. & Sadee, W. 2011. Intronic polymorphism in CYP3A4 affects hepatic expression and response to statin drugs. *The pharmacogenomics journal*, 11(4): 274-286.

Werk, A.N. & Cascorbi, I. 2014. Functional gene variants of CYP3A4. *Clinical Pharmacology & Therapeutics*, 96(3): 340-348.

Westerlind, K.C., Gibson, K.J., Evans, G.L. & Turner, R.T. 2000. The catechol estrogen, 4-hydroxyestrone, has tissue-specific estrogen actions. *Journal of endocrinology*, 167(2): 281-288.

Whiteford, N 2021. A Brief History of UV-Vis for Nucleic Acid Quantification. ASeq Newsletter, 15 Nov. https://aseq.substack.com/p/the-lunatic-from-unchained-labs Date of access: 20 Jun. 2023.

Wiggans, G.R., VanRaden, P.M., Bacheller, L.R., Tooker, M.E., Hutchison, J.L., Cooper, T. & Sonstegard, T.S. 2010. Selection and management of DNA markers for use in genomic evaluation. *Journal of dairy science*, 93(5): 2287-2292.

Williams, J.A. & Phillips, D.H. 2000. Mammary expression of xenobiotic metabolizing enzymes and their potential role in breast cancer. *Cancer research*, 60(17): 4667-4677.

Yager, J.D. 2012. Catechol-O-methyltransferase: characteristics, polymorphisms and role in breast cancer. *Drug Discovery Today: Disease Mechanisms*, 9(1-2): e41-e46.

Yekutieli, D. 2008. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481): 309-316.

Yuan, L. & Kaplowitz, N. 2009. Glutathione in liver diseases and hepatotoxicity. *Molecular aspects of medicine*, 30(1-2): 29-41.0

Zahid, M., Mondal, B., LeVan, T.D. & Rogan, E.G. 2018. Estrogen metabolism in African-American women with and without breast cancer: A pilot study. *Chemical research in toxicology*, 32(1): 190-194.

Zhang, K., Calabrese, P., Nordborg, M. & Sun, F. 2002. Haplotype block structure and its applications to association studies: power and study designs. *The American Journal of Human Genetics*, 71(6): 1386-1394.

Zhang, H., Cui, D., Wang, B., Han, Y.H., Balimane, P., Yang, Z., Sinz, M. & Rodrigues, A.D., 2007. Pharmacokinetic drug interactions involving 17α-ethinylestradiol: a new look at an old drug. *Clinical pharmacokinetics*, 46: 133-157.

Zhang, S., Shang, P., Gao, K., Zhao, G., Zhou, J., Chen, R., Ning, X. & Guo, C. 2022. Dynamics of estrogen-induced ROS and DNA strand break generation in estrogen receptor α-positive breast cancer. *Biochemical and Biophysical Research Communications*, 602: 170-178.

Zhao S, Jing W, Samuels DC, Sheng Q, Shyr Y, & Guo Y. 2018. Strategies for processing and quality control of Illumina genotyping arrays. *Briefings in bioinformatics*. 19(5):765-75.

Ziegler, R.G., Fuhrman, B.J., Moore, S.C. & Matthews, C.E. 2015. Epidemiologic studies of estrogen metabolism and breast cancer. *Steroids*, 99: 67-75.

Zou, K. & Ing, N.H. 1998. Oestradiol up-regulates oestrogen receptor, cyclophilin, and glyceraldehyde phosphate dehydrogenase mRNA concentrations in endometrium, but down-regulates them in liver. *The Journal of steroid biochemistry and molecular biology*, 64(5-6): 231-237.

SUPPLEMENTAL DATA

Table S0.1 74 Samples Spectrophotometry Range Results

Samples		Nanodrop		Qubit	Supplied Weight DNA (ng)	Samples		Nanodrop		Qubit	Supplied Weight DNA (ng)
Sample NO.	Sample ID	A_{260}/A_{280} Average	A_{260}/A_{230} Average	Concentration (ng/µl)	Concentration (≈ 50 ng/µl)	Sample NO.	Sample ID	A_{260}/A_{280} Average	A_{260}/A_{230} Average	Concentration (ng/µl)	Concentration (≈ 50 ng/µl)
1	eBOSS001	1.88	1.92	70.80	750	39	eBOSS045	1.85	1.89	53.20	750
2	eBOSS002	1.87	2.03	79.60	750	40	eBOSS046	1.89	1.93	56.40	750
3	eBOSS003	1.89	1.81	77.00	750	41	eBOSS047	1.94	1.92	73.60	751
4	eBOSS004	1.89	1.71	59.00	750	42	eBOSS049	1.81	1.49	59.20	752
5	eBOSS005	1.88	1.68	58.00	750	43	eBOSS050	1.86	1.57	37.00	740
6	eBOSS008	1.86	1.90	70.20	750	44	eBOSS052	1.86	1.63	59.00	755
7	eBOSS011	1.88	1.82	61.00	750	45	eBOSS053	1.84	1.92	95.00	750
8	eBOSS012	1.95	1.68	66.20	750	46	eBOSS054	1.81	1.81	91.80	753
9	eBOSS013	2.00	1.79	62.40	750	47	eBOSS055	1.87	1.36	20.60	680
10	eBOSS014	1.93	1.47	66.60	750	48	eBOSS056	1.89	1.47	22.80	684
11	eBOSS016	1.86	1.70	63.00	750	49	eBOSS057	1.84	1.45	34.40	722
12	eBOSS017	1.88	1.97	59.40	750	50	eBOSS058	1.87	1.58	39.80	756
13	eBOSS018	1.88	1.49	92.00	750	51	eBOSS060	1.83	1.58	58.20	751
14	eBOSS019	1.87	1.58	54.20	750	52	eBOSS061	1.88	1.38	43.60	741
15	eBOSS020	1.82	1.59	67.20	750	53	eBOSS064	1.81	1.84	112.00	753
16	eBOSS021	1.83	1.57	51.40	750	54	eBOSS067	1.82	1.46	47.60	762
17	eBOSS022	1.93	1.78	54.60	750	55	eBOSS068	1.86	1.58	49.40	790
18	eBOSS023	1.94	1.76	65.80	750	56	eBOSS070	1.83	1.79	50.60	754
19	eBOSS024	1.83	1.64	83.60	750	57	eBOSS071	1.82	1.89	50.00	750
20	eBOSS025	1.83	1.77	83.50	750	58	eBOSS072	1.9	1.34	57.00	752
21	eBOSS026	1.89	1.61	29.60	710	59	eBOSS073	1.89	1.86	94.80	758
22	eBOSS028	1.82	1.80	96.40	750	60	eBOSS074	1.9	1.78	71.20	755
23	eBOSS029	1.96	1.85	65.20	750	61	eBOSS075	1.85	1.52	114.00	752
24	eBOSS030	1.95	1.93	70.00	750	62	eBOSS076	1.87	1.49	65.80	750
25	eBOSS031	1.93	1.69	77.20	750	63	eBOSS079	1.88	1.29	59.00	755
26	eBOSS032	1.99	1.35	15.2	456	64	eBOSS080	1.88	1.38	54.80	751
27	eBOSS033	1.81	1.47	52.00	750	65	eBOSS082	1.86	1.51	59.60	751
28	eBOSS034	1.95	2.12	70.80	750	66	eBOSS083	1.86	1.98	77.80	755
29	eBOSS035	1.79	1.49	51.60	750	67	eBOSS084	1.89	1.82	52.60	752
30	eBOSS036	1.81	1.63	56.00	750	68	eBOSS087	1.88	1.9	110.00	759
31	eBOSS037	1.82	1.48	65.20	750	69	eBOSS091	1.91	1.97	52.00	755
32	eBOSS038	1.95	2.08	82.40	750	70	eBOSS093	1.92	1.9	70.20	753
33	eBOSS039	1.79	1.62	104.00	750	71	eBOSS094	1.87	1.51	87.20	759
34	eBOSS040	1.95	1.95	62.80	750	72	eBOSS096	1.87	1.73	93.60	758
35	eBOSS041	1.96	1.84	77.40	750	73	eBOSS097	1.88	1.71	86.00	757
36	eBOSS042	1.93	2.14	102.00	750	74	eBOSS098	1.83	1.85	69.00	752
37	eBOSS043	1.93	2.21	89.20	750		**Min**	**1.79**	**1.29**	**15.20**	**456**
38	eBOSS044	1.94	2.04	64.20	750	**Range**	**Max**	**2.00**	**2.21**	**114.00**	**790**

Legend

All 74 samples gDNA integrity quality control and amount are listed within this table. Nanodrop measured the possible purity of gDNA. The concentration measured the amount of gDNA in concentration and the last column contains the supplied weight of gDNA sent for genotyping. All samples were sent for genotyping, although all the purple highlighted samples sent were possibly either salt contaminated or might have to few gDNA for efficient genotyping

Figure S0.1 2% Agarose gels of gDNA in addition to a sample distribution illustration

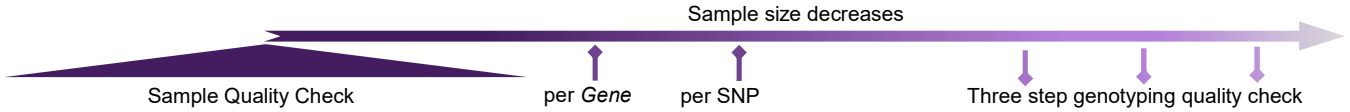
(a) **Illustration:** Gel 1 – 4, every filled circle and square were used as is in this study. Gel 1 – 3, the open squares were all selected and re – isolated and is shown in gel 4 and are now shown as filled shapes.

(b) **Gel 1** was a 50mℓ gel. Column 1 and 8 contained 17Kb positive control and column 2 was the negative control. Column 3-7 respectively shows sample NO. 1-5 (eBOSS001-005)

(c-e) **Gels 2-4** were 200mℓ gels with double combs were the pictures of each comb were combined at the grey stripe, dotted line to demonstrate the single gel. The first column in every row on every gel was a No limits 17Kb ladder. The second column of the top rows (2,4 and 6) of each gel were the negative controls. The last column of the top rows (2,4 and 6) of each gel as well as row 7 contained a 100bp ladder (SM0241). Row 2 (c) respectively showed sample NO. 6-22 (eBOSS008-028), row 3 (c) sample NO. 23-40 (eBOSS029-046), row 4 (d) sample NO. 41-57 (eBOSS047-071) and row 5 NO. 58-74 (eBOSS072-098). The last column of row 3 and 5 were empty wells.

Row 6 and 7 (e) showed all the repeated samples respectively in numerical order from sample NO.8 (eBOSS012) to sample NO.70 (eBOSS093). r^2

Table S0.2 74 Samples Summarized Allele Calls QC within GenomeStudio 2.0



Sample Name	Call Freq
eBOSS001	0.99
eBOSS002	0.99
eBOSS003	0.99
eBOSS004	0.99
eBOSS005	0.99
eBOSS008	0.99
eBOSS011	0.99
eBOSS012	0.99
eBOSS013	0.99
eBOSS014	0.99
eBOSS016	0.99
eBOSS017	0.99
eBOSS018	0.99
eBOSS019	0.99
eBOSS020	0.99
eBOSS021	0.99
eBOSS022	0.99
eBOSS023	0.99
eBOSS024	0.99
eBOSS025	0.99
eBOSS026	0.99
eBOSS028	0.99
eBOSS029	0.99
eBOSS030	0.99
eBOSS031	0.99
eBOSS032	0.99
eBOSS033	0.99
eBOSS034	0.99
eBOSS035	0.99
eBOSS036	0.99
eBOSS037	0.99
eBOSS038	0.99
eBOSS039	0.99
eBOSS040	0.99
eBOSS041	0.99
eBOSS042	0.99
eBOSS043	0.99
eBOSS044	0.99

Sample Name	Call Freq
eBOSS045	0.99
eBOSS046	0.99
eBOSS047	0.99
eBOSS049	0.99
eBOSS050	0.99
eBOSS052	0.83
eBOSS053	0.99
eBOSS054	0.99
eBOSS055	0.99
eBOSS056	0.99
eBOSS057	0.99
eBOSS058	0.99
eBOSS060	0.99
eBOSS061	0.99
eBOSS064	0.99
eBOSS067	0.99
eBOSS068	0.99
eBOSS070	0.99
eBOSS071	0.99
eBOSS072	0.99
eBOSS073	0.99
eBOSS074	0.99
eBOSS075	0.98
eBOSS076	0.98
eBOSS079	0.98
eBOSS080	0.99
eBOSS082	0.99
eBOSS083	0.99
eBOSS084	0.99
eBOSS087	0.99
eBOSS091	0.99
eBOSS093	0.99
eBOSS094	0.99
eBOSS096	0.99
eBOSS097	0.99
eBOSS098	0.98
min	0.83
max	0.99

Gene	SNPs	Sample Name	▲GC Score▲	GT Score	Cluster Sep
CYP1B1	rs1800440	All	0.81	0.79	1.00
	rs10012	eBOSS012	0.00		
		eBOSS070	0.00		
		eBOSS097	0.16	0.44	0.37
		eBOSS079	0.17	0.44	0.37
		eBOSS080	0.17	0.44	0.37
		eBOSS037	0.20	0.44	0.37
		eBOSS018	0.20	0.44	0.37
		eBOSS035	0.20	0.44	0.37
		eBOSS004	0.21	0.44	0.37
		eBOSS053	0.21	0.44	0.37
		eBOSS084	0.22	0.44	0.37
		eBOSS021	0.22	0.44	0.37
		The Rest	0.23	0.44	0.37
	rs2551188	All	0.80	0.80	1.00
CYP3A4	rs2246709	All	0.92	0.88	1.00
CYP1A2	rs2472304	All	0.94	0.90	1.00
	rs2470890	All	0.88	0.85	1.00
COMT	rs4646312	eBOSS067	0.84	0.86	0.96
		eBOSS031	0.85	0.86	0.96
		eBOSS094	0.88	0.86	0.96
		eBOSS016	0.89	0.86	0.96
		eBOSS047	0.89	0.86	0.96
		eBOSS049	0.89	0.86	0.96
		eBOSS038	0.89	0.86	0.96
		eBOSS012	0.89	0.86	0.96
		eBOSS033	0.90	0.86	0.96
		eBOSS028	0.90	0.86	0.96
		eBOSS017	0.90	0.86	0.96
		eBOSS003	0.90	0.86	0.96
		eBOSS046	0.90	0.86	0.96
		The Rest	0.90	0.86	0.96
	rs6269	All	0.82	0.80	1.00
	rs4633	eBOSS064	0.74	0.89	0.93
		eBOSS076	0.87	0.89	0.93
		The Rest	0.92	0.89	0.93
	rs2239393	All	0.80	0.79	0.98
	rs4818	eBOSS054	0.74	0.88	0.92
		eBOSS057	0.86	0.88	0.92
		The Rest	0.92	0.88	0.92
	rs4680	All	0.94	0.91	1.00
	rs165599	All	0.85	0.83	1.00

Legend

Call Freq - Call Frequency; *CYP* - Cytochrome P450; *COMT* - Catechol - O - methyltransferase;
GC Score - GenCall score; GT Score - GenTrain score; Cluster Sep - Cluster Separation score;
The Rest - Remainder of Samples Coloured / Highlighted sampleswere removed

The illustration above the table demonstrated the order of genotyping results and quality control. Per sample, all the 74 samples collective successful genotyping percentages is shown in the Call Freq columns. All purple highlighted samples within this table were removed from further association analysis. The genes were arranged according to chromosome number. The SNPs were arranged to the position on the chromosome. Only SNPs that significantly associated are listed within this table. The Sample Names were arranged according to an increasing GC core. The GC score is a parameter that determines whether a genotype can be accurately assigned to the specific sample. The GT score and Cluster Sep are parameters that determine proper genotype cluster formation and if the cluster are distinct from one another respectively.

Table S0.3 Chi-square test between ObsHET and PredHET for different populations

African					Legend
Gene	SNPs	Gene	SNPs		*CYP* - Cytochrome P450;
CYP3A4	rs4646437	***COMT***	rs933271	rs5993883	COMT - Catechol - O - methyltransferase;
ObsHET	4	ObsHET	7	6	African - African Control;
ObsHOM	19	ObsHOM	16	17	Caucasian - Caucasian Controls and COC users
PredHET	10	PredHET	11	11	COC - Combined contraceptive users (Caucasian)
PredHOM	13	PredHOM	12	12	Combined Controls - African Control and Caucasian Control
X^2	**0.02**	X^2	**0.08**	**0.04**	X^2 - Chi square test
Caucasian					All the significant chi - sqaure caculation are shown wihtin this table. The x2 was calculated by the ObsHET and PredHET and served as mutation interference QC parameter
None					
Combined Controls					
Gene	SNPs		Gene	SNPs	
CYP1B1	rs1056837	rs1056836	***CYP1A1***	rs2606345	
ObsHET	5	5	ObsHET	5	
ObsHOM	18	18	ObsHOM	18	
PredHET	11	11	PredHET	11	
PredHOM	12	12	PredHOM	12	
X^2	**0.02**	**0.01**	X^2	**0.02**	

Table S0.4 Caucasian Control group metabolite Shapiro-Wilk Normalization Results

Caucasian Controls	Descriptive Statistics									
	Unaltered					Tranformed - Log and Outliers Removed (3 x std. dev.)				
Metabolites	Maximum	Mean	Std. Deviation	Skewness	Kurtosis	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Estradiol_3_glucuronide	100	20	31	1.8	2.0	2.00	0.84	0.63	0.66	-0.81
Estradiol_17_sulfate	21	3	5	2.5	6.5	1.32	0.65	0.40	0.06	-0.87
Estradiol_3_sulfate	581	38	116	4.6	22.3	2.76	0.93	0.71	0.71	0.14
Estriol_3_sulfate	1016	79	208	4.2	18.6	3.01	1.23	0.70	0.91	0.34
Estriol_16_glucuronide	4144	434	828	4.1	18.2	3.62	2.24	0.59	0.30	-0.22
Estrone_3_sulfate	269	33	63	2.7	7.9	2.43	0.90	0.74	0.77	-0.85
Estrone_3_glucuronide	136347	9611	28221	4.2	18.5	5.13	3.00	0.87	0.85	0.31
16_Epiestriol	23	3	6	2.1	4.0	1.81	0.76	0.44	0.26	-0.11
16_Hydroxystrone	20	3	4	2.8	9.5	1.31	0.63	0.41	-0.08	-1.14
16_Ketoestradiol	16	3	4	2.6	6.3	1.21	0.54	0.36	0.57	-0.41
17_Epiestriol	109	14	31	2.4	4.7	2.04	0.77	0.62	0.92	-0.45
2_hydroxyestradiol	869	180	221	1.8	3.3	2.94	1.78	0.82	-0.54	-0.90
2_hydroxyestrone	188	43	54	1.5	1.3	2.28	1.25	0.64	-0.10	-0.98
2_Methoxyestradiol	1	0	0	4.0	16.6	1.96	1.25	0.45	-0.89	1.07
2_Methoxyestrone	17	1	3	4.4	20.8	1.75	0.78	0.49	0.22	-1.00
4_Hydroxyestradiol	56	13	15	1.5	2.1	1.75	0.95	0.50	-0.42	-0.77
4_Hydroxyestrone	146	24	31	2.8	9.6	2.16	1.05	0.59	-0.22	-0.84
4_Methoxyestradiol	1	0	0	3.7	15.7	2.13	1.21	0.46	-0.30	0.09
4_Methoxyestrone	127	7	25	4.8	23.7	1.47	0.02	1.92	-4.66	22.75
Androstenedione	153	21	31	3.7	15.3	2.19	1.13	0.38	0.97	0.92
Estradiol_Alpha	3	1	1	1.8	2.3	1.30	0.50	0.33	0.92	0.44
Estradiol_Beta	12	2	3	2.9	7.8	1.25	0.42	0.36	1.01	-0.14
Estriol	30	5	8	2.2	4.8	1.48	0.57	0.41	0.75	-0.62
Estrone	888	123	205	3.0	8.9	2.95	1.74	0.52	0.72	0.14
Estrone_2_hydroxy_3_methyl_ether	1	0	0	2.4	5.4	2.16	0.98	0.49	0.06	0.47
Progesterone	42	9	9	2.5	7.6	1.63	0.83	0.36	-0.03	0.23
Testosterone	191	11	38	4.9	24.3	1.23	0.00	1.91	-4.74	23.26
D_2OHE1_6N3A	380	49	88	3.1	9.4	2.58	1.33	0.52	0.84	0.49
D_2OHE2_6N3A	123	11	24	4.5	21.0	2.09	0.67	0.49	1.01	1.39
D_4OHE1_1N7G	7352	664	1517	3.9	16.9	3.87	1.86	1.09	0.02	-1.21
D_4OHE1_1N3A	9	2	2	2.1	4.2	0.97	0.02	1.90	-4.79	23.58
D_4OHE2_1N7G	3298	332	724	3.5	12.6	3.52	1.90	0.75	0.35	-0.41
D_4OHE2_1N3A	32	5	7	2.8	8.3	1.50	0.55	0.40	0.95	0.08
ROS	103	73	12	0.7	0.6	2.01	1.86	0.07	0.35	-0.15
GSHt	586	366	59	2.2	7.8	2.62	2.09	2.31	-5.00	24.98
FRAP	1273	951	225	-2.0	7.0	3.10	2.95	0.19	-4.10	18.86
Total deviated Metabolite distributions				**29**	**22**				**5**	**5**

Legend

4OHE1/2 - N (G / A) - estrogen DNA adduct; E1 - Estrone; E2 - Estradiol; ROS - Reactive oxygen species
GSHt - Glutathione transferase; FRAP - Ferric reducing ability of plasma

This table lists the metabolite transformation to induce normal distribution for the African population. There are 36 metabolites that are estrogen metabolite, conjugates and adducts as well as redox balance metabolites. There is unaltered and transformed data for each metabolite. The skewness and kurtosis are parameters for normal distribution determining a measure for symmetry and measurement of peak sharpness respectively. All purple highlighted samples are not acceptably normally distributed and were still used for association analysis.

Table S0.5 African Control group metabolite Shapiro-Wilk Normalization Results

African Controls	Descriptive Statistics									
	Unaltered					Tranformed - Log and Outliers Removed (3 x std. dev.)				
Metabolites	Maximum	Mean	Std. Deviation	Skewness	Kurtosis	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Estradiol_3_glucuronide	2669	130	554.0	4.8	22.9	1.93	0.34	2.11	-4.24	19.46
Estradiol_17_sulfate	133	8	27.5	4.7	22.2	1.15	0.13	2.02	-4.56	21.43
Estradiol_3_sulfate	131	16	30.7	2.9	9.1	2.12	0.74	0.61	0.93	-0.30
Estriol_3_sulfate	758	114	187.3	2.5	6.3	2.88	1.48	0.83	-0.18	-0.90
Estriol_16_glucuronide	6167	531	1312.1	4.0	17.0	3.79	2.17	0.67	0.20	1.09
Estrone_3_sulfate	797	69	170.0	3.9	16.7	2.90	1.08	0.82	0.59	-0.61
Estrone_3_glucuronide	1823237	80959	379810.1	4.8	23.0	3.91	2.34	2.56	-4.31	19.75
16_Epiestriol	42	5	9.4	3.4	12.8	1.63	0.59	0.44	0.60	-0.32
16_Hydroxystrone	45	6	11.4	2.7	6.9	1.65	0.63	0.44	0.84	0.11
16_Ketoestradiol	23	3	5.2	3.2	11.1	1.37	0.49	0.35	0.64	0.00
17_Epiestriol	394	27	81.1	4.6	21.6	2.60	0.76	0.67	1.03	0.86
2_hydroxyestradiol	3440	500	800.4	2.7	8.1	3.54	2.08	0.92	-0.48	-0.36
2_hydroxyestrone	9992	535	2075.2	4.7	22.3	3.05	0.90	2.29	-3.92	17.46
2_Methoxyestradiol	1	0	0.1	1.5	1.9	2.22	1.20	0.58	0.32	-0.69
2_Methoxyestrone	32	3	7.5	3.3	10.5	1.77	0.69	0.54	0.86	-0.62
4_Hydroxyestradiol	77	18	23.2	1.6	1.3	1.89	0.96	0.56	0.10	-0.92
4_Hydroxyestrone	799	71	172.1	3.8	15.8	2.90	1.19	0.74	0.57	-0.20
4_Methoxyestradiol	2	0	0.4	3.9	16.7	2.40	1.09	0.55	0.41	-0.21
4_Methoxyestrone	147	14	33.2	3.4	12.4	2.17	0.61	0.60	1.44	1.12
Androstenedione	194	24	47.1	3.0	8.9	2.29	0.95	0.58	0.69	0.16
Estradiol_Alpha	29	2	6.3	3.8	15.4	1.46	0.56	0.40	0.24	-0.42
Estradiol_Beta	20	3	4.6	3.0	9.9	1.30	0.49	0.40	0.80	-0.62
Estriol	1984	92	412.5	4.8	23.0	1.49	0.13	2.04	-4.41	20.51
Estrone	4266	462	1030.4	3.0	9.2	3.63	1.92	0.78	0.77	-0.20
Estrone_2_hydroxy_3_methyl_ether	3	0	0.6	3.0	10.7	1.57	0.81	0.48	-0.26	-1.33
Progesterone	85	13	18.2	3.2	11.5	1.93	0.88	0.42	0.66	0.86
Testosterone	22	5	5.9	1.9	3.2	1.35	0.60	0.34	0.77	-0.01
D_2OHE1_6N3A	288	38	66.6	3.1	9.8	2.46	1.20	0.57	0.21	0.59
D_2OHE2_6N3A	42	8	9.9	2.3	6.0	1.63	0.67	0.45	0.54	-0.83
D_4OHE1_1N7G	20623	1208	4384.9	4.4	19.6	4.31	1.56	1.09	0.82	0.67
D_4OHE1_1N3A	4	1	0.9	1.2	1.7	1.05	0.37	0.26	1.08	0.74
D_4OHE2_1N7G	3071	233	665.5	4.0	16.6	3.49	1.60	0.77	0.42	0.90
D_4OHE2_1N3A	37	5	8.3	3.2	10.7	1.56	0.52	0.38	1.18	1.62
ROS	118	91	14.4	-0.5	0.0	2.07	1.95	0.07	-0.87	0.54
GSHt	1273	1049	144.9	0.0	-1.2	3.10	3.02	0.06	-0.20	-1.12
FRAP	406	301	50.6	0.2	-0.6	2.61	2.47	0.07	-0.17	-0.57
Total deviated Metabolite distributions				**29**	**26**				**5**	5

Legend

4OHE1/2 - N (G / A) - estrogen DNA adduct; E1 - Estrone; E2 - Estradiol; ROS - Reactive oxygen species
GSHt - Glutathione transferase; FRAP - Ferric reducing ability of plasma

This table lists the metabolite transformation to induce normal distribution for the African population. There are 36 metabolites that are estrogen metabolite, conjugates and adducts as well as redox balance metabolites. There is unaltered and transformed data for each metabolite. The skewness and kurtosis are parameters for normal distribution determining a measure for symmetry and measurement of peak sharpness respectively. All purple highlighted samples are not acceptably normally distributed and were still used for association analysis.

Table S0.6 Combined Controls group metabolite Shapiro-Wilk Normalization

Combined Controls	Descriptive Statistics									
	Unaltered					Tranformed - Log and Outliers Removed (3 x std. dev.)				
Metabolites	Maximum	Mean	Std. Deviation	Skewness	Kurtosis	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Estradiol_3_glucuronide	2669	72	384	6.9	47.5	2.00	0.60	1.53	-5.34	33.99
Estradiol_17_sulfate	133	5	19	6.4	43.1	1.32	0.40	1.44	-6.20	41.27
Estradiol_3_sulfate	581	28	86	5.9	38.0	2.76	0.84	0.67	0.82	0.00
Estriol_3_sulfate	1016	96	197	3.3	12.0	3.01	1.35	0.77	0.32	-0.74
Estriol_16_glucuronide	6167	480	1076	4.2	19.1	3.79	2.20	0.62	0.21	0.45
Estrone_3_sulfate	797	51	126	4.8	27.1	2.90	0.99	0.78	0.67	-0.72
Estrone_3_glucuronide	1823237	43799	263112	6.9	47.4	5.13	2.69	1.89	-5.17	32.55
16_Epiestriol	42	4	8	3.3	13.1	1.81	0.68	0.44	0.40	-0.41
16_Hydroxystrone	45	4	9	3.4	12.8	1.65	0.63	0.42	0.39	-0.53
16_Ketoestradiol	23	3	5	2.9	8.5	1.37	0.51	0.35	0.58	-0.33
17_Epiestriol	394	20	60	5.5	33.3	2.60	0.77	0.64	0.95	0.12
2_hydroxyestradiol	3440	333	592	3.7	16.4	3.54	1.92	0.87	-0.42	-0.62
2_hydroxyestrone	9992	279	1442	6.8	46.6	3.05	1.09	1.64	-5.04	31.21
2_Methoxyestradiol	1	0	0	3.8	17.4	2.22	1.23	0.51	-0.12	-0.21
2_Methoxyestrone	32	2	6	4.0	16.7	1.77	0.74	0.51	0.53	-0.93
4_Hydroxyestradiol	77	15	19	1.7	2.3	1.89	0.95	0.52	-0.12	-0.86
4_Hydroxyestrone	799	46	122	5.4	31.9	2.90	1.12	0.66	0.35	-0.14
4_Methoxyestradiol	2	0	0	4.6	25.1	2.40	1.15	0.51	0.06	-0.30
4_Methoxyestrone	147	11	29	3.8	14.9	2.17	0.30	1.46	-5.59	36.49
Androstenedione	194	23	39	3.3	11.0	2.29	1.04	0.49	0.53	0.38
Estradiol_Alpha	29	2	4	5.5	32.4	1.46	0.53	0.36	0.54	-0.22
Estradiol_Beta	20	2	4	3.1	10.4	1.30	0.46	0.38	0.89	-0.46
Estriol	1984	46	286	6.9	47.9	1.49	0.36	1.44	-5.99	39.55
Estrone	4266	285	740	4.3	20.2	3.63	1.83	0.66	0.92	0.47
Estrone_2_hydroxy_3_methyl_ether	3	0	0	3.7	17.6	2.16	0.90	0.49	-0.06	-0.28
Progesterone	85	11	14	3.6	16.1	1.93	0.85	0.39	0.39	0.60
Testosterone	191	8	27	6.6	44.6	1.35	0.29	1.41	-6.25	41.87
D_2OHE1_6N3A	380	44	78	3.1	9.5	2.58	1.27	0.54	0.44	0.50
D_2OHE2_6N3A	123	9	19	5.1	30.2	2.09	0.67	0.47	0.80	0.38
D_4OHE1_1N7G	20623	925	3202	5.4	31.9	4.31	1.72	1.09	0.38	-0.62
D_4OHE1_1N3A	9	2	2	2.7	8.4	1.05	0.18	1.38	-6.48	43.90
D_4OHE2_1N7G	3298	284	691	3.6	12.6	3.52	1.75	0.77	0.33	0.06
D_4OHE2_1N3A	37	5	8	2.9	8.7	1.56	0.53	0.39	1.02	0.54
ROS	118	81	16	0.3	-0.8	2.07	1.90	0.09	-0.04	-0.96
GSHt	1273	693	361	0.3	-1.7	3.10	2.53	1.72	-6.72	46.04
FRAP	1273	640	367	0.2	-1.7	3.10	2.72	0.28	-0.17	-1.46
Total deviated Metabolite distributions				**32**	**32**				**8**	**8**

Legend

Combined Controls - African and Caucasian Controls; 4OHE1/2 - N (G / A) - estrogen DNA adduct; E1 - Estrone; E2 - Estradiol; ROS - Reactive oxygen species; GSHt - Glutathione transferase; FRAP - Ferric reducing ability of plasma

This table lists the metabolite transformation to induce normal distribution for the African population. There are 36 metabolites that are estrogen metabolite, conjugates and adducts as well as redox balance metabolites. There is unaltered and transformed data for each metabolite. The skewness and kurtosis are parameters for normal distribution determining a measure for symmetry and measurement of peak sharpness respectively. All purple highlighted samples are not acceptably normally distributed and were still used for association analysis.

Table S0.7 COC group metabolite Shapiro-Wilk Normalization Results

COC	Descriptive Statistics									
	Unaltered					Tranformed - Log and Outliers Removed (3 x std. dev.)				
Metabolites	Maximum	Mean	Std. Deviation	Skewness	Kurtosis	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Estradiol_3_glucuronide	46.4	6.1	9.2	3.8	16.8	1.67	0.68	0.37	0.59	0.93
Estradiol_17_sulfate	9.9	1.0	2.0	3.8	16.2	1.42	0.66	0.45	-0.05	-1.33
Estradiol_3_sulfate	29.3	7.2	8.5	1.6	1.9	1.47	0.71	0.42	0.22	-0.80
Estriol_3_sulfate	76.8	10.6	20.9	2.4	5.0	1.89	0.64	0.58	0.97	-0.27
Estriol_16_glucuronide	557.0	126.0	126.2	2.1	4.9	2.75	1.91	0.46	-0.90	2.53
Estrone_3_sulfate	151.3	15.1	34.4	3.4	11.4	2.18	0.77	0.56	0.93	0.55
Estrone_3_glucuronide	32645.9	3720.0	8598.6	2.8	7.1	4.51	2.65	0.93	0.31	0.49
16_Epiestriol	27.8	1.5	5.6	4.8	23.6	2.16	1.06	0.55	0.08	-0.44
16_Hydroxystrone	16.9	1.0	3.3	4.9	24.2	1.59	0.80	0.49	-0.04	-1.24
16_Ketoestradiol	17.0	1.1	3.4	4.7	23.1	1.96	0.79	0.56	0.43	-0.54
17_Epiestriol	58.8	5.6	14.8	3.3	9.8	1.77	0.57	0.54	1.33	0.74
2_hydroxyestradiol	2571.1	369.8	649.5	2.4	5.8	3.41	1.83	0.98	-0.21	-1.00
2_hydroxyestrone	464.6	61.0	109.6	2.7	7.6	2.67	1.28	0.68	0.39	-0.62
2_Methoxyestradiol	0.7	0.1	0.1	4.1	18.4	2.79	1.55	0.59	-0.40	0.20
2_Methoxyestrone	10.6	1.6	2.5	2.5	6.5	2.01	0.60	0.56	1.04	0.50
4_Hydroxyestradiol	68.2	12.5	18.0	1.8	2.9	1.83	0.86	0.61	0.03	-1.37
4_Hydroxyestrone	151.2	22.3	37.5	2.3	5.2	2.18	0.97	0.58	0.65	-0.44
4_Methoxyestradiol	0.2	0.0	0.1	2.0	3.8	2.49	1.61	0.52	-0.23	-0.56
4_Methoxyestrone	75.1	6.9	16.9	3.3	11.8	1.88	0.85	0.42	0.99	0.48
Androstenedione	25.7	7.3	6.0	1.4	2.3	1.41	0.73	0.36	-0.01	-0.83
Estradiol_Alpha	2.9	0.4	0.6	3.8	16.3	1.06	0.59	0.28	-0.60	-0.44
Estradiol_Beta	3.1	0.5	0.8	2.6	5.9	1.83	0.78	0.40	0.47	0.34
Estriol	20.0	1.3	4.0	4.8	23.1	1.30	0.62	0.34	0.27	-0.23
Estrone	94.9	12.9	23.0	3.2	9.5	1.98	0.83	0.43	1.17	2.33
Estrone_2_hydroxy_3_methyl_ether	0.7	0.1	0.2	2.5	5.7	2.44	1.18	0.50	0.07	1.12
Progesterone	28.8	5.4	6.6	2.4	6.2	1.46	0.57	0.37	0.93	0.10
Testosterone	18.1	3.1	4.1	2.4	6.6	1.26	0.55	0.29	0.78	-0.03
D_2OHE1_6N3A	56.6	12.2	13.6	1.9	3.8	1.75	0.86	0.46	0.17	-0.76
D_2OHE2_6N3A	40.1	4.2	7.8	4.4	20.4	0.98	0.07	1.91	-4.86	24.05
D_4OHE1_1N7G	2722.7	329.9	786.8	2.6	5.9	3.44	1.27	1.09	0.79	-0.59
D_4OHE1_1N3A	3.8	0.8	1.0	1.7	2.5	1.12	0.16	1.93	-4.80	23.65
D_4OHE2_1N7G	621.0	81.7	157.8	2.7	7.0	2.79	1.23	0.81	0.43	-0.89
D_4OHE2_1N3A	13.7	2.4	3.8	2.4	4.8	1.14	0.04	1.91	-4.80	23.65
ROS	211.8	164.5	24.8	0.1	-0.1	2.33	2.21	0.07	-0.32	0.21
GSHt	490.9	302.5	68.5	0.7	1.0	2.69	2.47	0.10	0.06	0.05
FRAP	1188.5	881.4	141.1	0.1	0.2	3.08	2.94	0.07	-0.35	0.36
Total deviated Metabolite distributions				**28**	**15**				**3**	**3**

Legend

COC - Combined oral contraceptive users (Caucasian); 4OHE1/2 - N (G / A) - estrogen DNA adduct; E1 - Estrone; E2 - Estradiol;
ROS - Reactive oxygen species; GSHt - Glutathione transferase; FRAP - Ferric reducing ability of plasma

This table lists the metabolite transformation to induce normal distribution for the African population. There are 36 metabolites that are estrogen metabolite, conjugates and adducts as well as redox balance metabolites. There is unaltered and transformed data for each metabolite. The skewness and kurtosis are parameters for normal distribution determining a measure for symmetry and measurement of peak sharpness respectively. All purple highlighted samples are not acceptably normally distributed and were still used for association analysis.

Table S0.8 Comparison of SPSS association analyses of different significance and metabolite distribution

Population group	Association	Distribution	Metabolite	SNP	*p*-values	
					GLM	U
African Control	Significant	Normal	ROS	Drs2246709	0.016	0.014
		Skew	2-HE_1	Drs4680	0.016	0.044
	Negligible	Normal	16-KetoE_2	Drs2687116	0.376	0.091
		Skew	2-HE_1	Drs4818	0.312	0.193
Caucasian Control	Significant	Normal	ROS	Rrs2551188	0.043	0.026
		Skew	None			
	Negligible	Normal	E_1-3-glucu	Drs10012	0.162	0.101
		Skew	D-4OHE_1-1N3A	Rrs2551188	0.116	0.072
Combined Controls	Significant	Normal	E_3-16-glucu	Drs4646312	0.008	0.005
		Skew	Testosterone	Drs2246709	0.018	0.015
	Negligible	Normal	FRAP	Drs740603	0.076	0.018
		Skew	4-MOE_1	Drs10916	< 0.001	< 0.001
COC	Significant	Normal	E_3-16-glucu	Drs1800440	0.045	0.026
		Skew	None			
	Negligible	Normal	2-MOE_1	Drs2606345	0.722	0.367
		Skew	D-4OHE2-1N3A	Drs10012	0.214	0.331

Legend

Combined Controls - African and Caucasian Controls; COC - Combined oral contraceptive users (Caucasian); 4OHE1/2-N (G / A) - oestrogen DNA adduct; E_1 - Oestrone; E_2 - Oestradiol; ROS - Reactive oxygen species FRAP - Ferric reducing ability of plasma; GLM - General linear model; U - Mann Whitney U (non - parametric) D-SNP – Dominant model; R-SNP – Recessive model

This table compares two different association analysis tests, GLM and Mann-Whitney U. GLM includes covariates with data that moderately deviated from normal distribution. U is more accurate for non - normal distributed data as well as small sample sizes. The Comparison were done per population group including significant, non - significant (negligible) associations and normal distributed and non - normal distributed (skew) metabolites. The purple highlighted p - values in the GLM column had significant covariate association. The green highlighted p - value in the U column had a significant value when GLM did not. The dominant or recessive models were used for calculation depending whether there was sufficient sample size and is indicated by a D or R at the beginning of the rsID.