

https://dspace.org/

A systematic assessment of the copy number variation (CNV) landscape in adme genes in Sub-Saharan African populations

Cottino, Laura Ann Alice

2022

https://hdl.handle.net/10539/34349

Downloaded from a DSpace repository. https://dspace.org/

A SYSTEMATIC ASSESSMENT OF THE COPY NUMBER VARIATION (CNV) LANDSCAPE IN ADME GENES IN SUB-SAHARAN AFRICAN POPULATIONS



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Laura Ann Alice Cottino

810316

A Thesis submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, 2022

DECLARATION

I, Laura Ann Alice Cottino (810316) declare that this Thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



08 June 2022

_____ _____

(Signature of candidate) (Date)

DEDICATION

This thesis is dedicated to my father, Cristian Cottino, who’s passion for acquiring knowledge has been a source of inspiration throughout my life and to my grandmother, Ann Walters, without whom my education would not have been possible.

PRESENTATIONS AND PUBLICATIONS ARISING FROM THIS RESEARCH PROJECT

Presentations:

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. “A Systematic Assessment of the CNV landscape in ADME genes in Sub-Saharan African Populations” (Poster). South African Society of Bioinformatics (SASBi)/South African Genetics Society (SAGS) Joint Conference & Student Symposium. Free State, South Africa. 2018 (16–18 October).

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. “A Systematic Assessment of the CNV landscape in ADME genes in Sub-Saharan African Populations” (Poster). 4th Global Genomic Medicine Collaborative (G2MC) Conference & Young Investigators Forum. Cape Town, South Africa. 2018 (28–30 November).

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. “A Systematic Assessment of the CNV landscape in ADME genes in Sub-Saharan African Populations” (Poster). Southern African Society for Human Genetics (SASHG) Biennial Congress. Cape Town, South Africa. 2019 (3–6 August).

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. “A Systematic Assessment of the CNV landscape in ADME genes in Sub-Saharan African Populations” (Oral). South African Society of Bioinformatics Student Council (SASBi-SC)/South African Society of Human Genetics (SASHG) Young Researchers' Symposium. Cape Town, South Africa. 2019 (3 August).

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. “A Systematic Assessment of the CNV landscape in ADME genes in Sub-Saharan African Populations” (Poster). Molecular Bioscience Research Thrust (MBRT) Postgraduate Research Day. Johannesburg, South Africa. 2019 (28 November).

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. An Assessment of the Copy Number Variation (CNV) Landscape in Sub-Saharan African Populations using Short-

and Long-read Sequencing (Poster). SMRT Leiden 2021 – Young Investigator Virtual Conference. Virtual Conference. 2021 (26–27 May).

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. An Assessment of the Copy Number Variation (CNV) Landscape in Sub-Saharan African Populations using Short- and Long-read Sequencing (Oral). Southern African Society for Human Genetics (SASHG) Young Researchers Online Symposium. Virtual Conference. 2021 (25–26 August).

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. An Assessment of the Copy Number Variation (CNV) Landscape in Sub-Saharan African Populations using Short- and Long-read Sequencing (Oral).18th Human Heredity and Health in Africa (H3Africa) Consortium Meeting. Virtual Conference. 2021 (30 August–3 September).

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. An Assessment of the Copy Number Variation (CNV) Landscape in Sub-Saharan African Populations using Short- and Long-read Sequencing (Oral). South African Society of Bioinformatics (SASBi)/South African Genetics Society (SAGS) 2021 Virtual Student Symposium. Virtual Conference. 2021 (14–16 September).

Publications:

Laura Cottino, Scott Hazelhurst & Venesa Sahibdeen. “A Systematic Assessment of the CNV landscape in ADME genes in Sub-Saharan African Populations”. (Advanced draft).

ABSTRACT

Various types of genetic variation exist within the human genome, ranging from single nucleotide changes to large-scale structural alterations. In recent years, several studies have identified and revealed the importance of a class of intermediate scale structural variation (SV), known as copy number variation (CNV). Despite its important impact on genetic diversity as well as disease, CNV remains widely understudied. The African continent carries a disproportionate burden of disease, and the treatments available are often ineffective or result in adverse drug reactions (ADRs). Genetic variation in absorption, distribution, metabolism, and excretion (ADME) genes has been shown to have a major impact on the pharmacokinetics, efficacy and safety of drugs. While single nucleotide variants in ADME genes have been well characterised, studies focused on the CNV landscape in ADME genes are lacking, especially in African populations. Additionally, CNVs can be identified using several approaches, including short-read and long-read whole genome sequencing (WGS) as well as from genotyping array data, with each approach posing its own limitations and advantages. This study aimed to (i) describe the CNV landscape in ADME genes in sub-Saharan African (SSA) populations, (ii) compare the CNV results from the short- and long-read sequencing data across the genome in order to assess whether short-read sequencing is adequate for accurate CNV calls and (iii) assess the utility of the Infinium H3Africa Consortium Array v2 to call CNVs. CNVs were called from the short-read, high-coverage, WGS of 953 individuals from across SSA using several CNV calling tools. A merged dataset was generated from the outputs of Genome STRiP, Manta, Delly, Lumpy and GATK that consisted of 362 ADME CNVs. The results of this study show that CNV is an additional and important source of genetic variability within ADME genes that has significant implications for drug response and precision therapy. There are, however, various limitations of SV calling from short-read sequencing data, resulting in generally low sensitivity and high false discovery rates, and long-read approaches have been shown to outperform short-read approaches. In this study, eight samples were selected for PacBio SMRT sequencing and a total of 42 041 CNVs were identified throughout the genome. When comparing the short- and long-read CNVs from the same eight individuals, ~50% of the short-read CNVs were concordant with the long-read data. Overall, the long-read sequencing approach allowed for the identification of a large proportion of additional CNVs and provided invaluable insights into the SSA CNV landscape. Sequencing of additional samples will further improve

our knowledge. Lastly, genotype data generated from over 10 000 South African individuals on the Infinium H3Africa Consortium Array v2 was used to call CNVs with PennCNV. A total of 7 312 CNVs were identified across the genome. This approach was able to call large, rare CNVs but called far fewer CNVs compared to the sequencing approaches discussed above. This lack of resolution limits the utility of the array in generating a comprehensive population-wide CNV call set. Additionally, the concordance between the short-read CNVs and array CNVs was very low, but was far higher (~50%) when compared to the long-read CNVs. Overall, this study has shown that CNV is an important source of genetic variation and has provided a more complete set of pharmacogenetically relevant variants in the African context. Additionally, it has highlighted the challenges of calling CNVs, with a diverse set of results being obtained from the different approaches.

ACKNOWLEDGEMENTS

This work is as a result of a great deal of support, guidance and assistance from several people and institutions whom I would like to acknowledge. I would like to express my gratitude to my supervisors Dr Venesa Sahibdeen and Professor Scott Hazelhurst for the guidance and insight that I have received throughout this project. My gratitude extends to GlaxoSmithKline (GSK) for funding this project as well as for the personal funding received. Thank you to all of the members of the Wits GSK-H3A ADME Team for sharing their knowledge and supporting me throughout this project, as well as to Professor Scott Hazelhurst and Dr Phelelani Mpangase for the computational and bioinformatics assistance. Thank you to Dr David Jakubosky for his substantial assistance with the Genome STRiP quality control. My greatest thanks goes to the participants and data collection teams from all of the data sources, without whom this research would not have been possible. Thank you to the staff at the Sydney Brenner Institute for Molecular Bioscience (SBIMB) Biobank staff for assisting with DNA extraction and preparation for sequencing, as well as to the SBIMB projects office for their assistance. I would like to offer my thanks to my fellow students at the SBIMB for their friendship, guidance and encouragement throughout the project. Finally, I would like to express my deepest gratitude towards my parents, Debbie and Cristian, my sisters, Cristina and Arianna, my grandmother Ann and my partner Grant for their unwavering support and belief in me.

TABLE OF CONTENTS

1. CHAPTER 1: LITERATURE REVIEW AND INTRODUCTION 1

1.1 Introduction 2

1.2 Pharmacogenomics and ADME genes 2

1.3 Copy Number Variation (CNV) 7

1.3.1 Types of CNVs 8

1.3.2 Effects of CNVs and distribution throughout the genome 9

1.3.3 Techniques used to identify CNVs 11

1.4 CNVs and Pharmacogenomics 17

1.4.1 CYP2D6 17

1.4.2 CYP2A6 19

1.4.3 Glutathione S-transferases (GSTs) 19

1.4.4 UDP-glucuronosyltransferases (UGTs) 20

1.4.5 SULT1A1 20

1.5 Important drugs and ADME genes in Sub-Saharan Africa (SSA) 21

1.5.1 Antiretrovirals (ARVs) 22

1.5.2 Anti-malarial drugs 24

1.5.3 Anti-TB drugs 25

1.5.4 Anticoagulants 26

1.5.5 Lipid-lowering drugs 27

1.5.6 Anti-cancer drugs 27

1.6 Pharmacogenomics in SSA 28

1.7 Rationale 31

1.8 Study Aim and Objectives 31
1.8.1 Aim 31
1.8.2 Objectives 31
1.9 Materials and Methods 31
1.10 Thesis outline 32
1.11 Reference list 33

2. CHAPTER 2: AN ASSESSMENT OF THE CNV LANDSCAPE IN ADME GENES USING SHORT-READ SEQUENCING APPROACHES. 46

2.1 Introduction 47
2.2 Materials and Methods 49
2.2.1 Study participants 49
2.2.2 Selection of ADME genes 51
2.2.3 CNV calling, processing and quality control 51
2.2.4 Merging of CNVs from different tools 62
2.2.5 CNV annotation 62
2.3 Results 63
2.3.1 Overview of CNVs identified by each tool 63
2.3.2 Overlap between callers 68
2.3.3 Merged CNV dataset 69
2.4 Discussion 83
2.4.1 Differences between CNV calling tools 83
2.4.2 Landscape of CNVs in ADME genes 84
2.4.3 Functional annotation and biological implications of CNVRs 88
2.4.4 Important ADME CNVs in an African context 90

2.4.5 Conclusions 97

2.5 Reference list 99

3. CHAPTER 3: VALIDATION OF SHORT-READ CNVS USING A LONG-READ SEQUENCING APPROACH 105

3.1 Introduction 106

3.2 Materials and Methods 110

3.2.1 Study participants 110

3.2.2 HiFi sequencing 110

3.2.3 Data analysis 111

3.3 Results 113

3.3.1 Sequencing quality 113

3.3.2 CNV landscape across the genome 115

3.3.3 Consequences of CNVs 119

3.3.4 Comparison to short-read CNVs 121

3.3.5 CNVs in ADME genes 123

3.4 Discussion 125

3.5 Reference list 129

4. CHAPTER 4: AN ASSESSMENT OF THE UTILITY OF THE INFINIUM H3AFRICA CONSORTIUM ARRAY V2 TO CALL CNVS 131

4.1 Introduction 132

4.2 Materials and Methods 134

4.2.1 Study participants 134

4.2.2 Data analysis 135

4.3 Results 138

4.3.1 Summary of array CNVs 138

4.3.2 Comparison to short-read data and long-read data 143

4.4 Discussion 145

4.5 Reference list 147

5. CHAPTER 5: OVERALL DISCUSSION AND CONCLUSION 149

6. REFERENCES 155

7. APPENDICES 171

APPENDIX A: PharmaADME core and extended ADME gene lists. 171

APPENDIX B: Depth of coverage plots for long-read sequences 179

APPENDIX C: Upset plots showing CNV intersections between samples of the same self-identified ethnolinguistic group. 189

APPENDIX D: Approval of title. 193

APPENDIX E: Ethical clearance certificate (M180738). 194

APPENDIX F: Ethical clearance certificate for use of AWI-Gen data (classified as a sub-study under M170880) (M2009102). 195

APPENDIX G: Plagiarism declaration. 197

APPENDIX H: Acknowledgment of Turnitin report by supervisors. 198

APPENDIX I: Turnitin report. 199

LIST OF FIGURES

CHAPTER 1

Figure 1.1: Hepatic Phase I and Phase II metabolizing enzymes and the reactions that they catalyse (adapted from Maheshwari et al., 2018). 4

Figure 1.2: Two major categories of CNVs (A) Recurrent CNV with a similar size and clustered/recurrent breakpoints. Recurrent CNVs are most commonly the result of non-allelic homologous recombination (NAHR) when there is a misalignment of the large flanking segmental duplications during meiosis. (B) Non-recurrent CNV with non-recurrent breakpoints occurring randomly throughout the genomic region of interest. The molecular mechanisms behind the formation of non-recurrent CNVs are not well understood (adapted from Gu, Zhang and Lupski, 2008). 9

Figure 1.3: Consequences of CNVs based on the gene element that is affected (adapted from Cabianca and Gabellini, 2010a). 10

Figure 1.4: Four main approaches used for the detection of CNVs from WGS data, namely (1) Read-Pair (RP); (2) Split-read (SR); (3) Read-depth (RD) and (4) Assembly (AS) (Pirooznia at al., 2015). 14

Figure 1.5: Overview of SV in the CYP2D6 gene. (A) Reference gene locus. (B) Deletion of entire CYP2D6 gene (CYP2D6*5). (C) Duplications or multiplications of functional or non-functional alleles. (D) SVs including CYP2D6*4. (E) SVs including CYP2D6*36 (Del Tredici et al., 2018). 18

CHAPTER 2

Figure 2.1: (A) Number of samples across west, central, east and southern Africa; (B) Number of samples per country. 50

Figure 2.2: Breakdown of the number of Phase I metabolising enzymes, Phase II metabolising enzymes, transporters and modifiers that make up the core and extended ADME gene lists. 51

Figure 2.3: CNV calling workflow including the approaches used by each tool to call CNVs. 953 BAM files were run individually through Manta, Delly, Lumpy and GATK and 919 BAM files were run jointly though Genome STRiP..... 52

Figure 2.4: Description of the 12 stages of Genome STRiP's CNVDiscovery pipeline. 53

Figure 2.5: SVE workflow (https://github.com/timothyjamesbecker/SVE). 58

Figure 2.6: Size ranges of ADME CNVs detected by Manta (blue), SURVIVOR (orange) and Genome STRiP (grey). 66

Figure 2.7: Allele frequency versus size for (A) all Genome STRiP CNVs; (B) all Manta CNVs; (C) all SURVIVOR CNVs; (D) GenomeSTRiP CNVs < 10kb; (E) Manta CNVs < 10kb and (F) SURVIVOR CNVs < 10kb. 67

Figure 2.8: Overlap between the CNVs called by Genome STRiP (green), Manta (blue) and SURVIVOR (Delly, Lumpy and GATK) (orange) for (A) all CNVs (n = 362) and (B) CNVs between 1kb and 100kb in size (n = 162). 68

Figure 2.9: Breakdown of the type of ADME gene affected by CNVs (Count; Percentage)..... 70

Figure 2.10: Number of CNVs per ADME gene. 71

Figure 2.11: Number of CNVs per ADME gene for Phase I metaboliser genes (n = 71) (green), Phase II metaboliser genes (n = 45) (blue), transporter genes (n = 41) (purple) and modifier genes (n = 12) (red). 72

Figure 2.12: Number of ADME CNVs per individual. 73

Figure 2.13: Total number of ADME CNVs identified in each region. The number of samples from each region is indicated in brackets below the region name. Note that the number of samples from each region are not equal. 74

Figure 2.14: Overlap between ADME CNVs identified in individuals from west (orange), central (green), east (blue) and southern (yellow) Africa. 75

Figure 2.15: Size range of ADME CNVs (A) All CNVs; (B) CNVs between 50 and 1 000bp in size; (C) CNVs between 1 000 and 10 000bp in size and (D) CNVs between 10 000bp and 100 000bp in size. 77

Figure 2.16: Size versus allele frequency for (A) all ADME CNVs; (B) ADME CNVs > 100 000bp excluded and (C) ADME CNVs > 10 000bp excluded. 78

Figure 2.17: Most severe consequence of each CNV obtained from Ensembl's VEP (Count; Percentage). 79

Figure 2.18: Breakdown of coding region affected by CNVs involving at least one exon. 80

CHAPTER 3

Figure 3.1: Comparison of short-read and Pacific Biosciences (PacBio) HiFi sequencing approaches (PacBio, 2020). 106

Figure 3.2: Flowchart of the PacBio HiFi sequencing approach (Wenger et al. 2019). 107

Figure 3.3: Number of deletions (blue) and duplications (orange) identified on each chromosome divided by chromosome length. 117

Figure 3.4: Upset plot showing the overlap between the CNVs identified in each sample. 118

Figure 3.5: Size distribution of deletions (blue) and duplications (orange) across the genome. 119

Figure 3.6: Most severe consequence of CNVs as predicted by Ensembl's VEP. "Other" includes splice region variants (n = 11); mature miRNA variants (n = 4); stop losses (n = 2) and start losses (n = 1).....120

Figure 3.7: Overlap between CNVs called from short-read data (A) Genome STRiP; (B) Manta; (C) SURVIVOR and the long-read sequencing data obtained from the same 8 individuals (based on 50% reciprocal overlap).....123

CHAPTER 4

Figure 4.1: Map showing the locations of participants in the H3Africa AWI-Gen study. 134

Figure 4.2: Distribution of CNVs across the chromosomes. 138

Figure 4.3: Size range of CNVs called from the array data. 141

Figure 4.4: Most severe consequence of CNVs as predicted by Ensembl's VEP. ..142

Figure 4.5: Overlap between CNVs called from the array data and short-read sequencing data (A) Genome STRiP; (B) Manta; (C) SURVIVOR from the same 100 South African individuals (based on 50% reciprocal overlap).....143

Figure 4.6: Overlap between CNVs called from the array data and long-read sequencing data from the same eight individuals (based on 50% reciprocal overlap). 144

LIST OF APPENDIX FIGURES

APPENDIX B

Figure B1: AB0330 depth of coverage.....179

Figure B2: AB2543 depth of coverage.....180

Figure B3: AJK0F depth of coverage.....181

Figure B4: CFF0X depth of coverage.....182

Figure B5: CFF0X (re-sequenced) depth of coverage.....183

Figure B6: CFF0X (combined) depth of coverage.....184

Figure B7: CVN0L depth of coverage.....185

Figure B8: CWF0F depth of coverage.....186

Figure B9: DEC0V depth of coverage. 187

Figure B10: DSX0S depth of coverage.....188

APPENDIX C

Figure C1: CNV overlap between CVN0L and DSX0S (Tswana samples).....189

Figure C2: CNV overlap between AB2543 and AB0330 (Tsonga samples). 190

Figure C3: CNV overlap between AJK0F and DEC0V (Venda samples).....192

Figure C4: CNV overlap between CFF0X and CWF0F (Xhosa samples).....192

LIST OF TABLES

CHAPTER 1

Table 1.1: Consensus terms for allele functional status and phenotype from the Clinical Pharmacogenetics Implementation Consortium (Caudle et al., 2017). 6

Table 1.2: Major and minor genes involved in the ADME of commonly used antiretrovirals, antimalarial and anti-TB drugs. 23

CHAPTER 2

Table 2.1: Source, country and self-identified ethnolinguistic group of sequences (n = 953) utilized in this study, as well as coverage and number of sequences. 49

Table 2.2: Count, length and frequency of ADME deletions, duplications and mCNVs identified using Genome STRiP, Manta as well as Delly, Lumpy and GATK (SURVIVOR dataset). 63

Table 2.3: Count, length and frequency of ADME deletions, duplications and mCNVs in the merged dataset. 69

Table 2.4: Drugs used to treat common diseases in Africa, genes that are involved in the ADME of these drugs as well as the CNVs identified in these genes (type, size, frequency and gene region affected by the CNV). 81

CHAPTER 3

Table 3.1: Various quality metrics for the HiFi sequencing of eight samples..... 114

Table 3.2: Number of CNVs and unique CNVs identified in each sample. 115

Table 3.3: Number of CNVs (deletions and duplications) on each chromosome..... 116

Table 3.4: Comparison between the short-read CNVs (Genome STRiP, Manta and SURVIVOR dataset) and long-read CNVs (pbsv) in terms of the number of CNVs detected, average number of CNVs detected per sample and the percentage of novel CNVs detected. 121

Table 3.5: List of core (n = 21) and extended (n = 80) ADME genes affected by CNVs called from the long-read data.....123

CHAPTER 4

Table 4.1: List of core (n = 21) and extended (n = 119) ADME affected by CNVs called from the Infinium H3Africa Consortium Array v2. 139

LIST OF APPENDIX TABLES

APPENDIX A

Table A1: Core ADME gene list (PharmaADME.org, 2021). 171

Table A2: Extended ADME gene list. The rankings of the extended ADME genes have been established based on input from seven major pharmaceutical companies in line with their interest in seeing a particular gene on a product (PharmaADME.org, 2021). 172

ABBREVIATIONS

1KGP	**:**	1000 Genomes Project
ABC	**:**	ATP-binding Cassette
ACT	**:**	Artemisin-based Combination Therapy
ADH	**:**	Alcohol Dehydrogenase
ADR	**:**	Adverse Drug Reaction
ADME	**:**	Absorption, Distribution, Metabolism and Excretion
AL	**:**	Artemether + Lumefantrine
ALDH	**:**	Aldehyde Dehydrogenase
ART	**:**	Antiretroviral Treatment
ARV	**:**	Antiretroviral
AS	**:**	Assembly
ASAQ	**:**	Artesunate + Amodiaquine
ASMQ	**:**	Artesunate + Mefloquine
BAF	**:**	B-Allele Frequency
CN	**:**	Copy Number
CNV	**:**	Copy Number Variation/Variant
CNVR	**:**	Copy Number Variation Region
CYP450	**:**	Cytochrome P-450
ddPCR	**:**	Digital Droplet PCR
DGV	**:**	Database of Genomic Variation
DRMAA	**:**	Distributed Resource Management Application API
FDA	**:**	Food and Drug Administration
FISH	**:**	Fluorescence *In Situ* Hybridisation
FDC	**:**	Fixed Dose Combination
GST	**:**	Glutathione S-transferase
GWAS	**:**	Genome-Wide Association Study/Studies
H3Africa	**:**	Human Heredity and Health in Africa
HGP	**:**	Human Genome Project
HiFi	**:**	High Fidelity
HIV	**:**	Human Immunodeficiency Virus
HMM	**:**	Hidden Markov Model
IIH	**:**	Isoniazid-Induced Hepatotoxicity
IM	**:**	Intermediate Metaboliser

INDEL	**:**	Insertion Deletion
kb	**:**	Kilobase/s
LD	**:**	Linkage Disequilibrium
LRR	**:**	Log R Ratio
LSF	**:**	Load Sharing Facility
LMIC	**:**	Low- to Middle-Income Country
mCNV	**:**	Multi-allelic CNV
Mb	**:**	Megabase/s
NAHR	**:**	Non-allelic Homologous Recombination
NAT	**:**	N-acetyltransferase
NCD	**:**	Non-Communicable Disease
NGS	**:**	Next Generation Sequencing
NM	**:**	Normal Metaboliser
NumCNV	**:**	Number of Called CNVs
ONT	**:**	Oxford Nanopore Technology
PacBio	**:**	Pacific Biosciences
PanGenIe	**:**	Pangenome-based Genome Inference
PD	**:**	Pharmacodynamics
P.f	**:**	*Plasmodium falciparum*
PFB	**:**	Population Frequency of the B Allele
PK	**:**	Pharmacokinetics
PM	**:**	Poor Metaboliser
RD	**:**	Read Depth
RM	**:**	Rapid Metaboliser
RP	**:**	Read Pair
SGE	**:**	Sun Grid Engine
SLC	**:**	Solute Carrier
SMRT	**:**	Single-Molecule Real-Time
SNP	**:**	Single Nucleotide Polymorphism
SNV	**:**	Single Nucleotide Variation/Variant
SR	**:**	Split Read
SSA	**:**	Sub-Saharan Africa/Sub-Saharan African
Strand-seq	**:**	Single-cell DNA Template Strand Sequencing
SULT	**:**	Sulfotransferase
SV	**:**	Structural Variation

TB	**:**	Tuberculosis
UGT	**:**	UDP-glucuronosyltransferase
UM	**:**	Ultrarapid Metaboliser
UTR	**:**	Untranslated Region
WES	**:**	Whole Exome Sequencing/Sequences
WF	**:**	Waviness Factor
WGS	**:**	Whole Genome Sequencing/Sequences
WHO	**:**	World Health Organization
ZMW	**:**	Zero-Mode Wave Guide

CHAPTER 1: LITERATURE REVIEW AND INTRODUCTION

1. CHAPTER 1: LITERATURE REVIEW AND INTRODUCTION

1.1 Introduction

The African continent is severely affected by disease – both communicable and non-communicable (Alessandrini and Pepper, 2014). Even though Africa accounts for only 14.5% of the global population, it carries ~30% of the global disease burden (Alessandrini and Pepper, 2014). With a high disease burden comes a large proportion of the population receiving some form of medication; however, it has become increasingly apparent that the majority of pharmaceutical drugs and their recommended doses do not have a uniform effect across all patients (Crews *et al.*, 2012). A given drug may be effective in one subset of patients, cause severe adverse drugs reactions (ADRs) in another, or may have no therapeutic effect at all, and an individual's genetic composition plays a major role in this variable response to drugs. Specifically, genetic variation in genes that encode proteins that play a role in the absorption, distribution, metabolism, and excretion (ADME) of drugs has been shown to have a major impact on the pharmacokinetics, efficacy and safety of drugs (Santos *et al.*, 2017). Genetic alterations that alter an individual's ability to absorb, distribute, metabolise and excrete drugs include single nucleotide variations (SNVs), insertion/deletion polymorphisms (INDELs) as well as copy number variations (CNVs) (Johansson and Ingelman-Sundberg, 2009). Currently, known SNVs only explain a portion of the variation in drug efficacy and toxicity that is observed (He, Hoskins and McLeod, 2011), and while SNVs have been extensively studied, a systematic assessment of CNV in ADME genes is lacking (Santos *et al.*, 2017). Whole genome sequencing (WGS) data will improve our knowledge of the CNV landscape in ADME genes. Furthermore, initiatives such as Human Heredity and Health in Africa (H3Africa) (Rotimi *et al.*, 2014) will aid in improving our knowledge in the African context, which likewise is understudied.

1.2 Pharmacogenomics and ADME genes

Every individual has a different genetic makeup that influences not only their susceptibility to disease, but also their response to various drugs (Ahmed *et al.*, 2016). While several factors contribute to an individual's response to a particular drug, including physiological conditions, environmental influences as well as pathological factors (Akhondzadeh, 2014), genetic factors have been found to account for a large

proportion of variability in drug response (Kalow, Tang and Endrenyi, 1998). These differences create challenges when it comes to the optimization of drug and dosage regimes (Wilkinson, 2005), as a given drug may work effectively in one group of patients but may be ineffective or result in serious ADRs, and even death in another group of patients (Ahmed *et al.*, 2016). Pharmacogenomics is therefore the study of how an individual's genetic variation alters their response to drugs, with the aim of using a patient's genotype as a rational means to optimise drug treatment (Saini, Saini and Sugandha, 2010; Crews *et al.*, 2012).

Heritability studies can provide an estimate of the relative contribution of genetic factors to a particular trait (Mroziewicz and Tyndale, 2010). It has been estimated that genetic factors account for 20–95% of the inter-individual variability in drug response that is observed, although this may vary from drug to drug (Kalow, Tang and Endrenyi, 1998). However, determining the heritability of a trait such as drug response using traditional twin or family studies is challenging as family members will rarely be exposed to the same drug in real life (Zhou and Pearson, 2013). Furthermore, heritability may vary across populations, and thus heritability calculations based on European populations may not be transferable to African populations (Zhu and Zhou, 2020).

Drug response is predominantly determined by the pharmacodynamic and pharmacokinetic properties of a particular drug, and genetic variation may alter the efficacy or toxicity of a drug by modifying either one or both of these properties (Ahmed *et al.*, 2016). Pharmacodynamics (PD) is defined as the study of how a drug affects its targets in a dose- and time-dependent fashion, while pharmacokinetics (PK) deals with how a drug is absorbed, distributed, metabolised and excreted over time (Pacey, Workman and Sarker, 2011). Absorption involves the movement of a drug from its site of delivery into the bloodstream (Paul, 2019). Once the drug has entered the bloodstream it is distributed throughout various tissues and fluids by drug transporters (Mehrotra *et al.*, 2007). Lastly, the elimination of the drug occurs via metabolism and excretion. Metabolism is the process whereby the drug is converted into molecules that are more easily excreted into urine or bile (Mehrotra *et al.*, 2007). Drug metabolism is predominantly carried out by Phase I and Phase II drug metabolising enzymes, and usually decreases the activity of the drug; however, in some cases the drug metabolites may have higher activity compared to the original compound (Mehrotra *et*

al., 2007). Excretion of the drug metabolites takes place via numerous pathways including the liver, lungs and skin, but the kidneys are the most significant organ for excretion (Mehrotra *et al.*, 2007).

Genetic variability in ADME genes explains, to a large extent, failure to respond to particular pharmaceutical drugs as well as ADRs when a standard dose is administered (Johansson and Ingelman-Sundberg, 2009). To date, PharmaADME has identified 32 core ADME genes (*Table A1*), which are known to influence drug ADME, as well as 267 extended ADME genes (*Table A2*) with less direct evidence (Hovelson *et al.*, 2017; PharmaADME.org, 2021). These genes are further separated into four categories: (1) Phase I metabolising enzymes that convert the parent drug into more active metabolites (*Figure 1.1)*; (2) Phase II metabolising enzymes that convert the parent drug into more inactive metabolites (*Figure 1.1)*; (3) transporters that are involved in the transport of the drug into and out of the cells and (4) modifiers that are able to either alter the gene expression of other ADME genes or the biochemistry of ADME enzymes (PharmaADME.org, 2021).

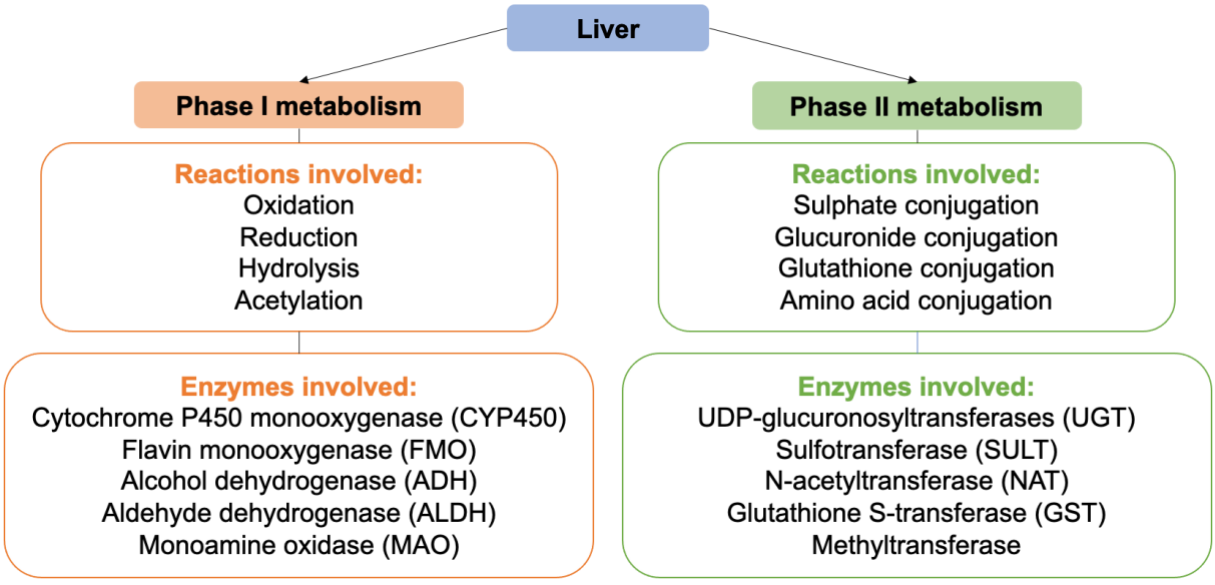


Figure 1.1: Hepatic Phase I and Phase II metabolising enzymes and the reactions that they catalyse (adapted from Maheshwari et al., 2018).

The cytochrome P-450 (CYP450) gene family is considered to be extremely important, if not the most important, for Phase I drug metabolism, based on the number of drugs that are metabolised by its protein products (*Figure 1.1)* (Mehrotra *et al.*, 2007). The *CYP3A* subfamily is responsible for the metabolism of 30–40% of all currently

prescribed drugs (Mehrotra *et al.*, 2007; Aka *et al.*, 2017), *CYP2D6* metabolises approximately 25% of drugs and the remaining CYPs metabolise a further 10% of drugs (Aka *et al.*, 2017). Other Phase I drug metabolising enzymes include the alcohol dehydrogenases (ADHs) and aldehyde dehydrogenases (ALDHs) (*Figure 1.1)* (Sarkar and Ganguly, 2016). The main Phase II drug metabolising enzymes include UDP-glucuronosyltransferases (UGTs), N-acetyltransferases (NATs), sulfotransferases (SULTs) and glutathione S-transferases (GSTs) (*Figure 1.1)* (Stingl, Brockmöller and Viviani, 2013).

Genetic variation within drug metabolising genes can result in several phenotypes which have been standardised by the Clinical Pharmacogenetics Implementation Consortium (Caudle *et al.*, 2017) as the following: (1) poor metaboliser (PM) phenotypes that are characterised by little to no enzyme activity due to combinations of non-functional alleles and/or decreased function alleles; (2) intermediate metaboliser (IM) phenotypes that are characterised by decreased enzyme activity due to combinations of normal function, decreased function, and/or non-functional alleles; (3) normal metaboliser (NM) phenotypes that are characterised by fully functional enzyme activity due to combinations of normal function and decreased function alleles; (4) rapid metaboliser (RM) phenotypes that are characterised by increased enzyme activity compared to NMs, but less than ultrarapid metabolisers, due to combinations of normal function and increased function alleles; and (5) ultrarapid metaboliser (UM) phenotypes that are characterised by increased enzyme activity compared to RMs due to two increased function alleles or more than two copies of normal function alleles (*Table 1.1)* (Stingl, Brockmöller and Viviani, 2013; Caudle *et al.*, 2017).

Drug transporters are also clinically important as they can modulate the pharmacological activity of drugs by affecting their intracellular concentrations (Scherrmann, 2009). The two main superfamilies of drug transporters include the ATP-binding cassettes (ABCs) and solute carriers (SLCs) (Scherrmann, 2009; Rajman, Knapp and Hanna, 2020). Genetic variation within these genes may contribute to inter-individual variability in drug response (Rajman, Knapp and Hanna, 2020) and four phenotypes have been standardized as: (1) poor function in which there is little to no transporter activity, (2) decreased function which is categorised by a decreased transporter function, (3) normal function where there is fully functional enzyme activity,

and (4) increased function which is categorised by increased transporter function compared to normal function (*Table 1.1)* (Caudle *et al.*, 2017).

Table 1.1: Consensus terms for allele functional status and phenotype from the Clinical Pharmacogenetics Implementation Consortium. (Caudle *et al.*, 2017, licensed under Creative Commons CC BY 4.0 https://creativecommons.org/licenses/by/4.0/. No changes have been made to the figure content.)

Term/gene category	Final term[a]	Functional definition	Genetic definition	Example diplotypes/alleles
Allele functional status: all genes	Increased function	Function greater than normal function	N/A	*CYP2C19*17*
	Normal function	Fully functional/wild-type	N/A	*CYP2C19*1*
	Decreased function	Function less than normal function	N/A	*CYP2C19*9*
	No function	Nonfunctional	N/A	*CYP2C19*2*
	Unknown function	No literature describing function or the allele is novel	N/A	*CYP2C19*29*
	Uncertain function	Literature supporting function is conflicting or weak	N/A	*CYP2C19*12*
Phenotype: drug-metabolizing enzymes (CYP2C19, CYP2D6, CYP3A5, CYP2C9, TPMT, DPYD, UGT1A1)	Ultrarapid metabolizer	Increased enzyme activity compared to rapid metabolizers	Two increased function alleles, or more than 2 normal function alleles	*CYP2C19*17/*17* *CYP2D6*1/*1XN*
	Rapid metabolizer	Increased enzyme activity compared to normal metabolizers but less than ultrarapid metabolizers	Combinations of normal function and increased function alleles	*CYP2C19*1/*17*
	Normal metabolizer	Fully functional enzyme activity	Combinations of normal function and decreased function alleles	*CYP2C19*1/*1*
	Intermediate metabolizer	Decreased enzyme activity (activity between normal and poor metabolizer)	Combinations of normal function, decreased function, and/or no function alleles	*CYP2C19*1/*2*
	Poor metabolizer	Little to no enzyme activity	Combination of no function alleles and/or decreased function alleles	*CYP2C19*2/*2*
Phenotype: transporters (SLCO1B1)	Increased function	Increased transporter function compared to normal function.	One or more increased function alleles	*SLCO1B1*1/*14*
	Normal function	Fully functional transporter function	Combinations of normal function and/or decreased function alleles	*SLCO1B1*1/*1*
	Decreased function	Decreased transporter function (function between normal and poor function)	Combinations of normal function, decreased function, and/or no function alleles	*SLCO1B1*1/*5*
	Poor function	Little to no transporter function	Combination of no function alleles and/or decreased function alleles	*SLCO1B1*5/*5*
Phenotype: high-risk genotype status (*HLA-B*)	Positive	Detection of high-risk allele	Homozygous or heterozygous for high-risk allele	*HLA-B*15:02*
	Negative	High-risk allele not detected	No copies of high-risk allele	

[a]All terms should begin with the gene name (e.g., CYP2D6 Poor metabolizer, TPMT Normal metabolizer, SLCO1B1 decreased function).

Clinically, the loss or reduction of enzyme activity can result in an altered response to drugs (Jarvis, Peter and Shaman, 2019). PMs and IMs may present with ADRs when a standard dose is administered due to the inability to clear the drug at a normal rate and toxic levels of the drug accumulating in the body (Jarvis, Peter and Shaman, 2019). These patients may either require lower doses of the drug, or an alternate drug. In the case of prodrugs (molecules with little to no pharmacological activity until converted to the active drug in the body) (Rautio *et al.*, 2018), PMs and IMs may fail to activate the prodrug to the levels required for effective treatment (Jarvis, Peter and Shaman, 2019). Increased enzyme activity also has important clinical implications. Standard doses of drugs administered in their active form may not produce effective

results in RMs and UMs as a result of the increased metabolism of the active drug into its excretable forms (Jarvis, Peter and Shaman, 2019). These patients may require increased doses in order to achieve effective treatment (Jarvis, Peter and Shaman, 2019). Conversely, for prodrugs, a lowered dose may be required for RMs and UMs as the increase in metabolism of the inactive prodrug into active metabolites may result in toxic levels of the active metabolite accumulating in the body (Jarvis, Peter and Shaman, 2019).

The use of pharmacogenomics to understand how genetic variation alters drug response will ultimately assist healthcare professionals in prescribing the most effective treatment for patients in terms of safety and efficacy (Ahmed *et al.*, 2016). It has been estimated that pharmacogenomics will decrease ADRs by 10-20% and increase the efficacy of treatment by 10–15% (Ingelman-Sundberg, 2004). This approach will ensure maximum efficacy while minimizing adverse effects, thereby decreasing the overall cost of treatment (Saini, Saini and Sugandha, 2010; Crews *et al.*, 2012). While pharmacogenetically relevant SNVs have been well characterised, an assessment of the CNV landscape is lacking.

1.3 Copy Number Variation (CNV)

Various types of genetic variation exist in the human genome, ranging from single nucleotide changes to large chromosomal alterations (Eichler, 2019). In recent years, numerous studies have uncovered an abundance of submicroscopic CNVs involving DNA fragments that range from kilobases (kb) to megabases (Mb) in size (Freeman *et al.*, 2006; Redon *et al.*, 2006). These deletions, duplications and complex multisite variants are known collectively as copy number variations, and have been found in all humans and various other species (Feuk, Carson and Scherer, 2006; Freeman *et al.*, 2006). CNVs can be defined as a type of structural variation (SV) with copy number alterations involving a fragment of DNA that is greater than 50 bp in size (Eichler, 2019). The lower size limit (50 bp) is an arbitrary threshold. Earlier studies defined the size of a CNV as larger than 1000 bp, followed by 100 bp, and currently, 50 bp (Pös *et al*., 2021). CNVs can be simple in structure, such as tandem deletions or duplications, or may be far more complex and involve gains or losses of homologous sequences at multiple sites across the genome (Redon *et al.*, 2006). Alleles containing between zero and 13 gene copies have been reported across the human genome (Weinshilboum, 2003).

One of the important insights gained by the Human Genome Project (HGP) was that the abundance of single nucleotide polymorphisms (SNPs) present across the genome was recognised as a major source of genetic variation (Ionita-laza *et al.*, 2010). This led to the idea that the majority of the phenotypic variability observed in the human population was the result of these single base changes (Freeman *et al.*, 2006). Consequently, major efforts were put into the detection and analysis of SNPs and little attention was paid to any other form of genetic variation (Ionita-laza *et al.*, 2010). Only recently, has the crucial role that CNVs play been recognised (Ionita-laza *et al.*, 2010), and it is now believed that CNVs contribute as much to genetic variation as SNPs do (Fiegler *et al.*, 2006; Freeman *et al.*, 2006). Furthermore, CNVs account for a large portion of the human genome, with approximately 12% of the human genome subjected to CNV (Hastings et al., 2009), further validating the importance of CNVs. Several databases containing known CNVs now exist and include the Database of Genomic Variants (DGV) which contains ~7 million CNVs (MacDonald *et al.*, 2014), the Children's Hospital of Philadelphia's Copy Number Variation Project which contains ~50 000 CNVs (Shaikh *et al.*, 2009) and various others.

1.3.1 Types of CNVs

Whilst little is known about the molecular mechanisms behind the formation of CNVs, they can be separated into two major classes, namely, recurrent CNVs and non-recurrent CNVs, which have distinctly different methods of formation as well as structures (Hastings, Ira and Lupski, 2009; Arlt, Wilson and Glover, 2012). Recurrent CNVs, which account for between 20% and 40% of CNVs, have recurrent or common breakpoints that occur within segmental duplications, while non-recurrent CNVs have unique breakpoints that are scattered randomly across the genomic region and are not dependent on segmental duplications (*Figure 1.2)* (Gu, Zhang and Lupski, 2008; Hastings, Ira and Lupski, 2009; Arlt, Wilson and Glover, 2012). While most non-recurrent CNVs are tandem duplications or simple deletions, some are far more complex and may contain deleted and duplication regions within the same CNV, as well as pieces of normal sequences or inversions (Arlt, Wilson and Glover, 2012).

Recurrent CNVs are believed to arise due to non-allelic homologous recombination (NAHR), which is usually caused by the misalignment of large flanking segmental duplications during the process of meiosis (*Figure 1.2)* (De Smith *et al.*, 2009;

Hastings, Ira and Lupski, 2009; Arlt, Wilson and Glover, 2012). While the mechanism behind the formation of recurrent CNVs is fairly well understood, the molecular processes that result in non-recurrent CNVs have yet to be fully elucidated. It is suspected, however, that errors occurring in DNA replication during mitosis, rather than meiotic homologous recombination, give rise to non-recurrent CNVs (Arlt, Wilson and Glover, 2012). Secondly, similar CNVs have been induced experimentally through the inhibition of replication and the creation of "replication stress" conditions (Arlt, Wilson and Glover, 2012). This work suggests that the simplest model for the formation of non-recurrent CNVs involves the incorrect restoration of stalled replication forks during DNA replication (Arlt, Wilson and Glover, 2012). Further experimental work is required to better understand the mechanisms behind the formation of non-recurrent CNVs.

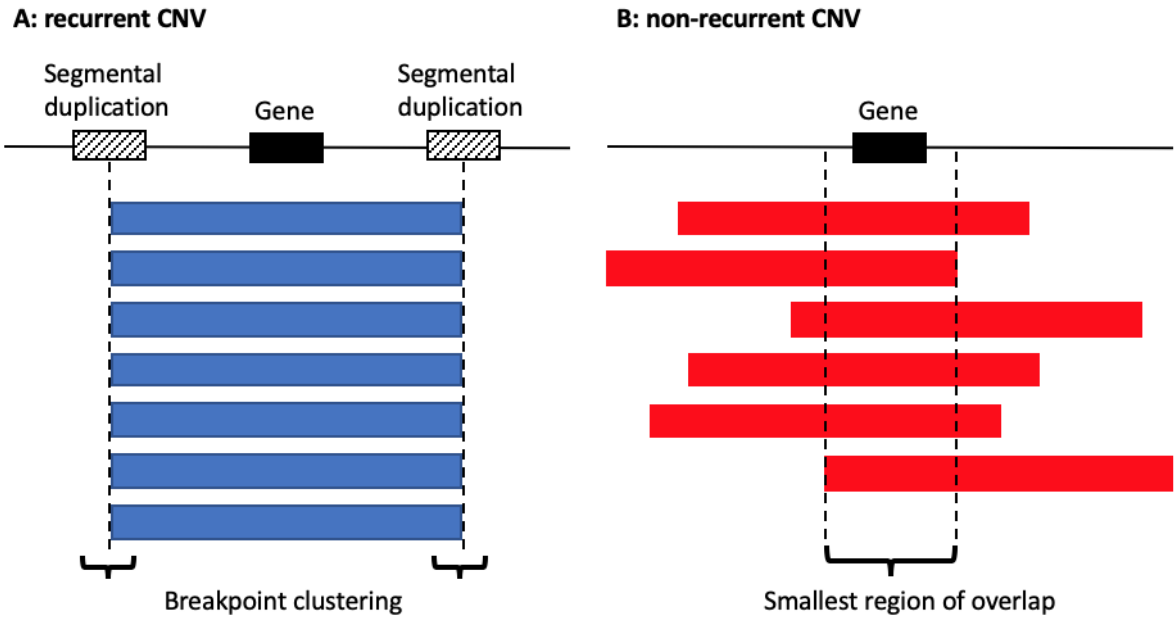


Figure 1.2: Two major categories of CNVs (A) Recurrent CNV with a similar size and clustered/recurrent breakpoints. Recurrent CNVs are most commonly the result of non-allelic homologous recombination (NAHR) when there is a misalignment of the large flanking segmental duplications during meiosis. (B) Non-recurrent CNV with non-recurrent breakpoints occurring randomly throughout the genomic region of interest (adapted from Gu, Zhang and Lupski, 2008).

1.3.2 Effects of CNVs and distribution throughout the genome

CNVs can have an effect on the phenotype by directly affecting gene dosage, thus altering the level of gene expression, and may also have significant impacts on protein

sequence, structure and function (Lee and Jeon, 2008; Zhao *et al.*, 2013). Depending on the gene element that is affected, CNVs can have varying consequences (Cabianca and Gabellini, 2010a). If an entire gene(s) locus is involved, the gene dosage may be altered resulting in either gene amplification or deletion (*Figure 1.3)* (Stranger *et al.*, 2007; De Smith *et al.*, 2009). Alternatively, if only part of a gene, such as an exon of a protein-coding gene, is affected by a CNV, the protein structure and function can be altered resulting in an aberrant protein being produced (*Figure 1.3)* (De Smith *et al.*, 2009; Cabianca and Gabellini, 2010a). CNVs can also lead to altered splicing or an alteration in the production of non-coding RNA (ncRNA) in cases where a non-coding portion of a gene is affected by a CNV (*Figure 1.3)* (Stranger *et al.*, 2007; De Smith *et al.*, 2009; Cabianca and Gabellini, 2010a). Lastly, CNVs within extragenic regions can alter gene expression if they occur within regulatory regions (*Figure 1.3)* (De Smith *et al.*, 2009). However, while copy number is known to be functionally significant, these functions have yet to be fully elucidated (Redon *et al.*, 2006).

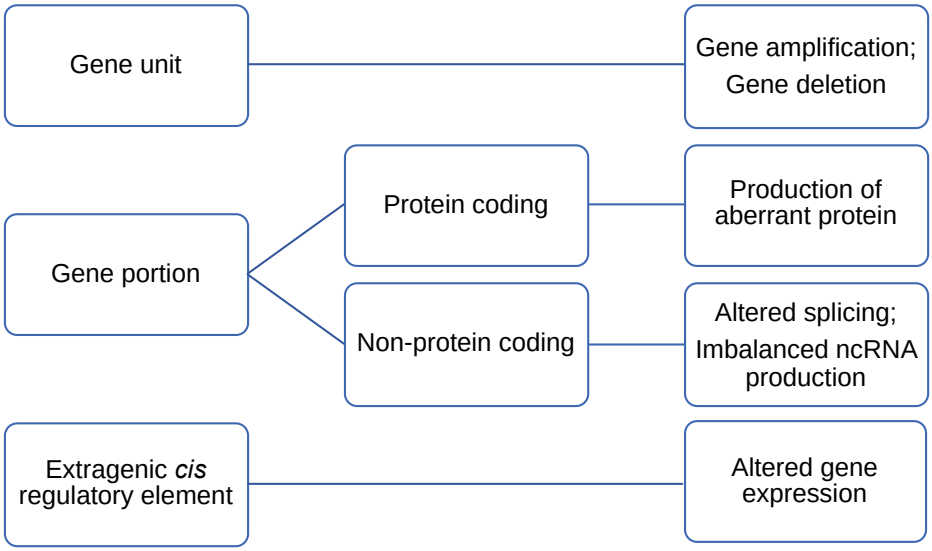


Figure 1.3: Consequences of CNVs based on the gene element that is affected (adapted from Cabianca and Gabellini, 2010a).

Studies have shown that CNVs are distributed heterogeneously/non-randomly across the genome (Redon *et al.*, 2006), and are significantly more likely to occur close to telomeres and centromeres due to the repetitive nature of these regions (Nguyen, Webber and Ponting, 2006; De Smith *et al.*, 2009). However, there is no region of the genome in which CNVs do not occur (Redon *et al.*, 2006). Interestingly, the relationship between the size of a chromosome and the number of CNVs it contains is non-linear

(De Smith *et al.*, 2009). Differences in genomic structure between the chromosomes plays a role in the number of CNVs a chromosome contains (De Smith *et al.*, 2009). For example, chromosome 18 has ~20% CNV sequence whereas chromosome 19 has greater than 40% CNV sequence (De Smith *et al.*, 2009). Chromosome 19 contains a large proportion of segmental duplications which may result in an increased presence of CNVs on this chromosome compared to other larger chromosomes (Grimwood *et al.*, 2004; De Smith *et al.*, 2009).

Furthermore, certain gene categories are enriched for CNVs (De Smith *et al.*, 2009). Gene Ontology analyses have revealed that genes involved in the immune system, development of the brain, neurophysiological processes and processing of external stimuli have been shown to be enriched for CNVs (Feuk, Carson and Scherer, 2006; de Smith *et al.*, 2007). On the other hand, genes that are involved in early development and cell signalling and division, as well as those involved in nucleic acid metabolism, have been shown to contain fewer CNVs (Conrad *et al.*, 2006; Redon *et al.*, 2006).

1.3.3 Techniques used to identify CNVs

Array-based CNV detection

CNVs were traditionally identified through cytogenetic techniques, including karyotyping and fluorescence *in situ* hybridisation (FISH), and, more recently, through array-based technologies, which allowed for a more comprehensive view of CNVs throughout the genome (Carter, 2007; Buysse *et al.*, 2009). High-throughput, high-density genotyping technologies that are used in genome-wide association studies (GWAS) can also be used for the detection of CNVs (Lin, Naj and Wang, 2013). Although genotyping arrays were originally developed in order to determine the bases present at certain SNPs, the signal intensity information derived from these arrays has increasingly been used to detect CNVs (Carter, 2007). In general, tools that can detect CNVs from array data are categorised as either individual-wise analysis or joint analysis of multiple individuals (Zhang *et al.*, 2019).

Tools that involve individual-wise analysis include PennCNV (Colella *et al.*, 2007) and QuantiSNP (Wang *et al.*, 2007). These approaches make use of various types of information obtained from SNP probes using Illumina platforms as well as other information such as population frequencies at each locus and linkage disequilibrium

(LD) structure between adjacent loci (Wang *et al.*, 2009; Zhang *et al.*, 2019). In general, individual-wise approaches are successful in detecting rare, large CNVs spanning at least ten probes on a SNP array (Zhang *et al.*, 2019). However, this type of analysis can be error-prone as these methods are reliant on the assumption that allele intensities are properly normalized such that they are comparable across probes throughout the genome and are less affected by differences such as GC content at different loci (Zhang *et al.*, 2019). Moreover, comparing CNVs between different individuals can be difficult, as the individual-level CNV calls are often uniform across individuals (Zhang *et al.*, 2019).

Joint analysis of multiple individuals takes advantage of consensus CNV signals across individuals, which are particularly useful for the detection of CNVs that occur in multiple individuals (Zhang *et al.*, 2019). Tools that utilise this approach include iPattern (Pinto *et al.*, 2011), Piet (Zhang, Lange and Sabatti, 2012), and msscan (Zhang *et al.*, 2019). These tools align CNVs that are found in multiple individuals into CNV regions (CNVRs), which makes downstream analysis easier (Zhang *et al.*, 2019). However, the exact breakpoints of the CNVs within CNVRs are often unknown and these approaches are less sensitive to detecting rare CNVs (Zhang et al., 2019).

Additionally, SNP haplotypes that tag or are able to predict the presence of a CNV can be identified (Nyangiri *et al.*, 2020). These haplotypes are not only associated with the presence of absence of a CNV, but also with the likely copy number (CN) of the CNV (Nyangiri *et al.*, 2020). Using this approach CNVs that have previously been tagged by SNP haplotypes can be inferred from SNP array data without having to perform CNV calling.

Short-read sequencing-based CNV detection

The recent emergence of next generation sequencing (NGS) technology such as short-read WGS has allowed for higher coverage and resolution of the genome as well as encouraged the development of tools for detecting and interpreting CNVs (Meyerson, Gabriel and Getz, 2010; Alkan, Coe and Eichler, 2011; Koboldt *et al.*, 2012; Zhao *et al.*, 2013; Santos *et al.*, 2017). This has allowed for a more accurate determination of CN, higher sensitivity with respect to the sizes and types of variation that can be detected, more accurate breakpoint determination and, lastly, an increased ability to detect novel CNVs (Meyerson, Gabriel and Getz, 2010; Alkan, Coe and Eichler, 2011;

Koboldt *et al.*, 2012). Importantly, NGS has allowed for the precise determination of breakpoints, which is critical in determining the functional impact of the variant as well as inferring the mutational mechanisms of origin (Mills *et al.*, 2011; Koboldt *et al.*, 2012). Currently, deep-coverage, paired-end WGS is the gold standard for CNV detection in population-wide studies and provides the highest sensitivity and resolution in CNV detection from short-read sequencing data (Zhou *et al.*, 2017).

There are four main approaches that can be used for the identification of CNVs from short-read WGS data, namely Read Pair (RP), Read Depth (RD), Split Read (SR) and Assembly (AS) methods (*Figure 1.4)* (Zhao *et al.*, 2013; Pirooznia, Goes and Zandi, 2015). The RP method uses read pairs whose distances are significantly different from the predetermined average insert size in order to identify CNVs (*Figure 1.4)* (Zhao *et al.*, 2013). In other words, the average insert size of the actual sequenced read-pairs is compared to the expected insert size based on the reference genome (Pirooznia, Goes and Zandi, 2015). An insert size that is shorter than predicted is an indication of a deletion, whereas an insert size that is longer than predicted is an indication of a duplication (Pirooznia, Goes and Zandi, 2015) (*Figure 1.4)*. However, this approach has a limited ability to call CNVs that are larger than the average insert size (Zhao *et al.*, 2013). The SR-based approach uses reads from paired-end sequencing where only one of the reads of the pair has mapped accurately while the other read has partially or completely failed to map to the reference genome (*Figure 1.4)* (Zhang *et al.*, 2011; Zhao *et al.*, 2013). The partially mapped reads or unmapped reads are then fragmented and the first and last fragments are aligned to the reference genome independently (Zhao *et al.*, 2013). This method is thus able to provide accurate breakpoint information, down to the single base pair level (Zhao *et al.*, 2013); however, it is limited to the length of the reads and thus cannot call larger CNVs (Pirooznia, Goes and Zandi, 2015). Both the RP and SP approaches are able to report the position of a CNV, but not the exact CN (Pirooznia, Goes and Zandi, 2015). RP-based tools include PEMer (Korbel *et al.*, 2009) and BreakDancer (Chen *et al.*, 2009), while Pindel (Ye *et al.*, 2009), SVseq2 (Zhang, Wang and Wu, 2012) and Prism make use of the SR approach include (Jiang, Wang and Brudno, 2012).

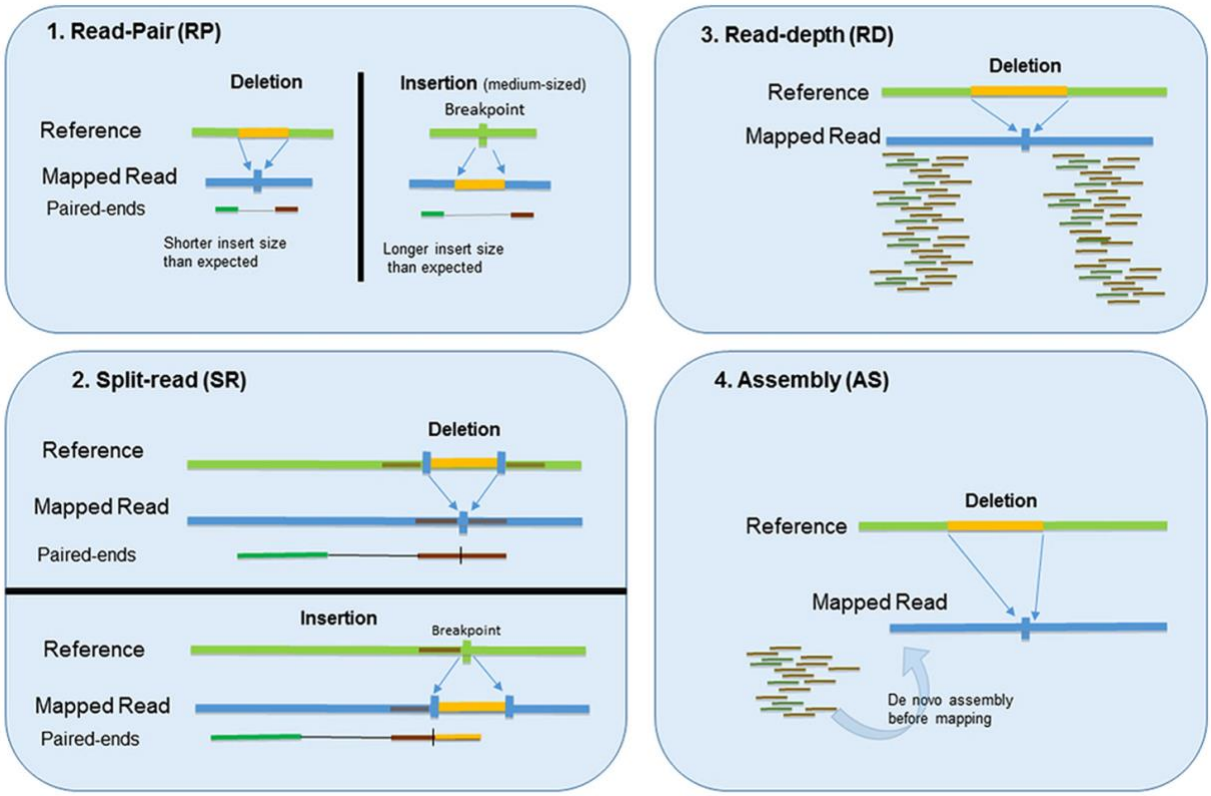


Figure 1.4: Four main approaches used for the detection of CNVs from WGS data, namely (1) Read-Pair (RP); (2) Split-read (SR); (3) Read-depth (RD) and (4) Assembly (AS). (Pirooznia at al., 2015, licensed under Creative Commons CC BY 4.0 https://creativecommons.org/licenses/by/4.0/. No changes have been made to the figure content.)

The RD-based approach is based on the hypothesis that depth of coverage is correlated with CN (*Figure 1.4)* (Teo *et al.*, 2012). Therefore, a deletion would result in a lower read depth than expected, while a duplication would result in a higher read depth than expected (Zhao *et al.*, 2013). In general, RD-based methods follow a four-step procedure: (1) mapping; the reads are aligned to the reference genome and the RD is determined by counting the number of reads within a predefined window; (2) normalisation; the read counts are normalized to account for potential biases such as GC content and repetitive regions; (3) estimation of CN; CN is estimated along the chromosome to identify gains and losses, and (4) segmentation; genomic regions with a similar CN are merged to detect discordant CN regions (Zhao *et al.*, 2013). Compared to the SR- and RP-based approaches, which can only determine the location of CNVs, RD-based methods are able to detect the exact CN (Zhao *et al.*, 2013; Pirooznia, Goes and Zandi, 2015). RD-based methods are also able to detect large CNVs and CNVs in complex genomic regions (Yoon *et al.*, 2009), but are unable

to determine exact breakpoints (Pirooznia, Goes and Zandi, 2015). CNVnator (Abyzov *et al.*, 2011), RDXplorer (Yoon *et al.*, 2009) and cn.MOPS (Klambauer *et al.*, 2012) make use of the RD-based approach.

All three of the above methods initially align the NGS reads to a reference genome before CNV detection (Zhao *et al.*, 2013). Instead, the AS approach first assembles overlapping reads to form contigs which are then compared to the reference genome in order to identify regions with inconsistent CN (*Figure 1.4)* (Nijkamp *et al.*, 2012; Teo *et al.*, 2012; Zhao *et al.*, 2013). This is an unbiased approach to detecting novel CNVs; however, this approach requires extensive computational power and performs poorly in complex genomic regions (Zhao *et al.*, 2013; Pirooznia, Goes and Zandi, 2015). Magnolya (Nijkamp *et al.*, 2012) makes use of the AS-based approach.

As discussed above, each approach has its own limitations and advantages. Therefore, more recent tools have combined different approaches in order to take advantage of the useful features of each approach, thereby facilitating more accurate CNV detection (Zhao *et al.*, 2013; Pirooznia, Goes and Zandi, 2015). Tools that make use of a combination of approaches include GenomeSTRiP (RP and RD) (Handsaker *et al.*, 2015), Manta (RP, SR and AS) (Chen *et al*., 2016), Lumpy (RP, SR and RD) (Layer *et al*., 2014) and Delly (RP and SR) (Rausch *et al*., 2012).

Long read sequencing-based CNV detection

Short-read sequencing (~150–300 bp read length) poses a major limitation in the identification of structural variants including CNVs. It has been demonstrated that even with the use of sophisticated algorithms, it remains difficult to accurately align short reads that originate from regions that contain SVs, thus making SV calling difficult and inaccurate (Mantere, Kersten and Hoischen, 2019). Furthermore, phasing information, which allows for more accurate population genetic analyses, is often lost with short-read sequencing data (Delaneau *et al.*, 2013), and breakpoints, that are important for determining the functional impact of a CNV, are difficult to determine (Mahmoud *et al.*, 2019). Additionally, the data analysis of short-read sequencing is highly dependent on a reference genome, which has been shown to introduce bias (Genovese *et al.*, 2013). This is particularly problematic for the detection of SVs as they have been shown to be highly population- or individual-specific (Chaisson, Wilson and Eichler, 2015).

The emergence of long-read sequencing technologies has the potential to overcome the above-mentioned issues associated with short-read sequencing, and provide an improvement in the characterization of genetic variants, including CNVs (Mantere, Kersten and Hoischen, 2019). With a read-length of >10 kb, longer reads offer the advantage of spanning an entire CNV, making alignment and CNV calling far more accurate (Mantere, Kersten and Hoischen, 2019). This approach also allows for phasing and the more accurate determination of CNV breakpoints. Importantly, long-read sequencing allows for *de novo* assembly of the genome, thereby avoiding the limitations associated with aligning reads to a reference genome (Chaisson, Wilson and Eichler, 2015).

The two long-read sequencing methods that are commercially available are (1) nanopore sequencing by Oxford Nanopore Technologies (ONT) and (2) single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio) (Clarke *et al.*, 2009; Eid *et al.*, 2009). Both PacBio and ONT sequencing methods can generate reads that are thousands of base pairs in length, thus spanning entire CNVs (Mahmoud *et al*., 2019). Current long-read based approaches are mostly always dependant on the underlying sequencing technology, with tools that can only be used on either PacBio data or ONT data, but not both (Mahmoud *et al*., 2019). For PacBio long-read data, three main tools for SV calling have been developed, namely PacBio's own tool, pbsv (Pacific Biosciences, 2017), as well as PBHoney (English, Salerno and Reid, 2014) and SMRT-SV (Huddleston *et al.*, 2017). For ONT data, NanoSV (Cretu Stancu *et al.*, 2017) has been developed for calling SVs. The only tool that can be used on both PacBio and ONT data is Sniffles (Sedlazeck *et al.*, 2018).

Even though we are still in the early stages of developing tools to call SVs from long-read sequencing data, overall, long-read approaches for calling SVs, including CNVs, have been seen to outperform short-read based approaches (Mahmoud *et al.*, 2019). Although the single-base error rate of this sequencing approach was initially high (Mahmoud *et al.*, 2019), it has been significantly reduced in recent years, with current error rates of <1% (Amarasinghe *et al.*, 2020). However, the requirement for relatively large quantities of high quality DNA (Ardui *et al.*, 2018) and the high cost of sequencing, has resulted in this approach not being used regularly for SV detection (Mahmoud *et al.*, 2019).

1.4 CNVs and Pharmacogenomics

Despite the fact that SNVs are the most widely studied source of genetic variation, they only explain a portion of the variation in drug efficacy and toxicity that is observed (Ramírez *et al.*, 2019). Recent studies have revealed that some of this unexplained variability in drug response may be due to the presence of CNVs (Redon *et al.*, 2006; Ramírez *et al.*, 2019). Functionally important CNVs within several important ADME genes, including *CYP2D6*, *CYP2A6* (Phase I metabolisers), *GSTM1*, *GSTT1*, *UGT2B17, UGT2B28* and *SULT1A1* (Phase II metabolisers), have already been identified (He, Hoskins and McLeod, 2011; Ramírez *et al.*, 2019; Tremmel *et al.*, 2020).

1.4.1 CYP2D6

The CYP2D6 Phase I drug metabolising enzyme is responsible for the metabolism of approximately 25% of prescribed drugs (Zhou, 2009; He, Hoskins and McLeod, 2011). However, the *CYP2D6* gene locus is highly polymorphic, making it one of the most difficult ADME genes to genotype (Goh *et al.*, 2017). *CYP2D6* is located on chromosome 22q13.1 and forms a gene cluster with pseudogenes, *CYP2D7* and *CYP2D8* (Goh *et al.*, 2017). The genetic complexities associated with *CYP2D6* are the result of SNVs within the gene, duplications and deletions of the gene as well as within the gene, and recombination events with the highly homologous neighbouring pseudogene, *CYP2D7* (Goh *et al.*, 2017; Del Tredici *et al.*, 2018). As a result of the large amounts of variation that exist within this gene, the function of *CYP2D6* is highly variable and phenotypes can range from PMs with no enzyme activity at all, to UMs with significantly increased enzyme activity (Del Tredici *et al.*, 2018).

CNVs within *CYP2D6* contribute significantly to the CYP2D6 phenotype diversity that is observed in the clinic (He, Hoskins and McLeod, 2011). The presence of more than two copies of *CYP2D6* normal functioning alleles (e.g.*1×N, *2×N, *33×N, *35×N whereby 2<N<13) (*Figure 1.5C*) (He, Hoskins and McLeod, 2011; Del Tredici *et al.*, 2018) has been shown to elevate CYP2D6 enzyme activity, and individuals with these alleles are classified as UMs of CYP2D6 substrates (Ingelman-Sundberg, 2005). In “Asian” (i.e. Central Asian, East Asian, South Asian, Southeast Asian or Western Asian) and “Caucasian” (i.e. European, American or similar) populations, duplicated normal functioning alleles are observed at a frequency of 0.5–1.5% (Sistonen *et al.*, 2007), while in African populations gene duplications with UM phenotypes are more

common, with a frequency between 1.6–3.3% (Griese *et al.*, 1999; Fuselli *et al.*, 2010). Multiple copies of non-functional alleles (*4×N, *36×N) (*Figure 1.5D*, *Figure 1.5E*) and reduced function alleles (*10×N, *17×N) (He, Hoskins and McLeod, 2011; Del Tredici *et al.*, 2018) have also been reported; however, these individuals do not display increased metabolism of CYP2D6 substrates (Ishiguro *et al.*, 2004). Gene deletion of one or both copies of *CYP2D6* (*5) (*Figure 1.5B*) (He, Hoskins and McLeod, 2011; Del Tredici *et al.*, 2018) results in the loss of production of CYP2D6 protein (Steen *et al.*, 1995). These individuals have impaired metabolism of CYP2D6 substrates and are classified as PMs (He et al., 2012). The frequency of *CYP2D6*5* gene deletions is fairly consistent across most populations (2–7%) (Bradford, 2002).

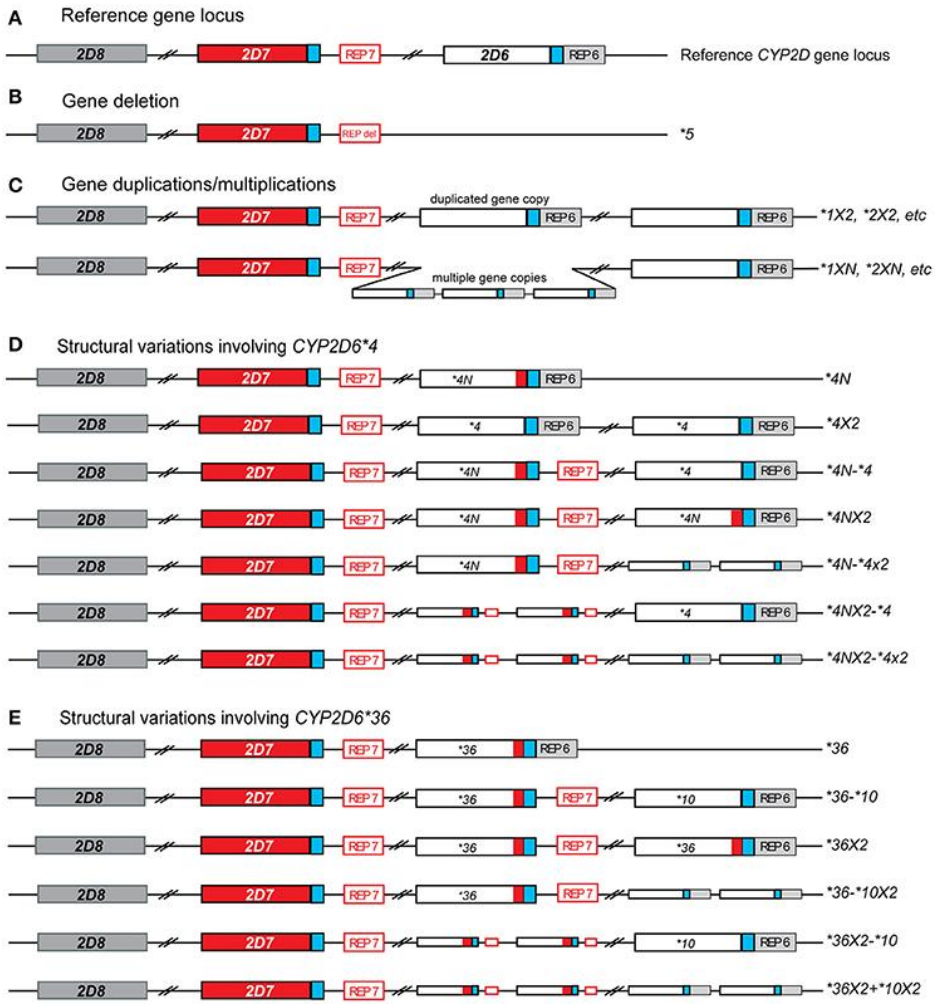


Figure 1.5: Overview of SV in the *CYP2D6* gene. (A) Reference gene locus. (B) Deletion of entire *CYP2D6* gene (*CYP2D6*5*). (C) Duplications or multiplications of functional or non-functional alleles. (D) SVs involving *CYP2D6*4*. (E) SVs involving *CYP2D6*36*. (Del Tredici *et al.*, 2018, licensed under Creative Commons CC BY 4.0 https://creativecommons.org/licenses/by/4.0/. No changes have been made to the figure content.)

1.4.2 CYP2A6

The CYP2A6 enzyme is primarily responsible for the metabolism of nicotine into its inactive metabolite cotinine (Tyndale and Sellers, 2001). Genetic variation within this gene has been shown to influence inter-individual variability in nicotine metabolism as well as smoking behaviour and cancer risk (Kamataki *et al.*, 2005). Both full gene deletions and duplications of the *CYP2A6* gene have been reported across various populations (Martis *et al.*, 2013). The full gene deletion allele is known as *CYP2D6*4,* and individuals who are homozygous for this deletion have a complete lack of enzyme activity (Ito *et al.*, 2015). The frequency of the *CYP2A6*4* allele varies across populations and ranges from 0% in Ashkenazi Jewish individuals to 1.5% in "Caucasian" (i.e. European, America or similar) individuals, 3.8% in African American individuals and 8.3% in "Asian" (i.e. Central Asian, East Asian, South Asian, Southeast Asian or Western Asian) individuals (Martis *et al.*, 2013). Full gene duplications of *CYP2A6* are less common and are generally observed at a frequency of < 2% across populations (Martis *et al.*, 2013).

1.4.3 Glutathione S-transferases (GSTs)

The GST enzyme superfamily is responsible for the detoxification of several exogenous substances including carcinogens and therapeutic drugs (He, Hoskins and McLeod, 2011). Two of the GST genes which have been shown to harbour CNVs are *GSTM1* and *GSTT1*, and studies have shown that full gene deletions of these genes are relatively common (He, Hoskins and McLeod, 2011). Approximately 50% and 30% of European individuals are homozygous for *GSTM1* and *GSTT1* gene deletions, respectively (Garte *et al.*, 2001; Piacentini *et al.*, 2011). These frequencies are slightly different in African populations with >27% of individuals being homozygous for *GSTM1* gene deletions and >37% of individuals being homozygous for *GSTT1* gene deletions (Garte *et al.*, 2001; Piacentini *et al.*, 2011). While *GSTM1* and *GSTT1* full gene deletions are common across various populations, *GSTT1* gene duplications have not been reported and *GSTM1* gene duplications are rare (He, Hoskins and McLeod, 2011). Due to their role in the detoxification of drugs, individuals with *GSTM1* and *GSTT1* genes deletions are at an increased risk of developing drug-related toxicities (He, Hoskins and McLeod, 2011).

1.4.4 UDP-glucuronosyltransferases (UGTs)

UGTs are the second largest contributor to drug metabolism after the CYP450 enzymes (Williams *et al.*, 2004). CNVs have been reported in two of the UGT genes, namely *UGT2B17* and *UGT2B28,* which have been identified as two of the most commonly deleted genes in the human genome (McCarroll *et al.*, 2006; Ménard *et al.*, 2009). *UGT2B17* is a Phase II metabolising enzyme and contributes to the metabolism of several drugs (Gaedigk and Leeder, 2012). Full gene deletions of the gene have been observed at a high frequency of ~60% in European individuals (Tremmel *et al.*, 2020), and approximately 12% of European individuals harbour a homozygous *UGT2B17* deletion (Wilson *et al.*, 2004; Gallagher *et al.*, 2007). *UGT2B17* gene duplications are extremely rare and have only been reported in a handful of individuals (Gaedigk and Leeder, 2012). Previous studies have shown that full gene deletions of *UGT2B28* are less common when compared to *UGT2B17* (Habibi *et al.*, 2017). In individuals of European descent, the *UGT2B28* gene deletion is observed at a frequency of 28.7% (Tremmel *et al.*, 2020). CNVs in these UGT genes seem to be a major source of variability in gene expression, and therefore may play a role in the variability in clinical response to various drugs (Ménard *et al.*, 2009).

1.4.5 SULT1A1

SULT1A1 is involved in the metabolism of various exogenous substances and its enzyme activity has an influence on the drug response to various drugs (Yu *et al.*, 2013). In fact, studies have shown that *SULT1A1* CNVs are highly correlated with enzyme activity, and that variability in enzyme activity is best explained by CNVs rather than SNPs (Hebbring *et al.*, 2007; Gaedigk and Leeder, 2012). Individuals with deletions or duplications are classified as slow sulfators and rapid sulfators respectively (Hebbring *et al.*, 2007). *SULT1A1* gene deletions and the resulting slow sulfator phenotype have been observed across multiple populations, with frequencies ranging from 3.3% in African American individuals to 8.4% in “Caucasian” individuals (i.e. European, American or similar) (Vijzelaar *et al.*, 2018). *SULT1A1* gene duplications, with a CN as high as six copies, are far more common and are observed at a frequency of 60.9% in African American individuals (Gaedigk and Leeder, 2012; Vijzelaar *et al.*, 2018). Additionally, individuals of African descent are more likely to have higher CNs (Hebbring *et al.*, 2007; Gaedigk and Leeder, 2012).

A recent study by Santos and colleagues provided the first assessment of CNV across ADME genes (Santos *et al.*, 2018). Their work involved the analysis of whole genome sequences (obtained from the 1000 Genomes Project (1KGP)) as well as whole exome sequences (obtained from the Exome Aggregation Consortium) in order to identify CNVs in 208 pharmacogenes (Santos *et al.*, 2018). Their study revealed that 97% of the ADME genes included in the study contained novel CNVs and in total, 2 611 deletions and 2 978 duplications were detected (Santos *et al.*, 2018). Deletions were identified in 84% of ADME genes, and all of the previously well-characterised CNVs involving *CYP2A6, CYP2D6, UGT2B17, UGT2B28, GSTM1, GSTT1* and *SULT1A1* were identified (Santos *et al.*, 2018). Novel ADME gene deletions were observed most commonly in African individuals at a frequency of 4.2%, and aggregated CNV frequencies were highest in African individuals (7.2%) (Santos *et al.*, 2018).

Tremmel and colleagues performed the most recent CNV profiling in pharmacogenes using pane-based exome resequencing (Tremmel *et al.*, 2020). Their study focussed on full-gene CNVs (encompassing at least 60% of the gene) and they identified 445 deletions and 167 duplications in 36 pharmacogenes. These included all of the well-known CNVs of *CYPs, GSTs, SULTs* and *UGTs*, previously described as rare CNVs, as well as novel CNVs (Tremmel *et al.*, 2020). The most frequent CNVs were deletions of *GSTM1, GSTT1* and *UGT2B17* (Tremmel *et al.*, 2020).

1.5 Important drugs and ADME genes in Sub-Saharan Africa (SSA)

Africa is the continent most affected by disease – both communicable and non-communicable (Alessandrini and Pepper, 2014). Although Africa accounts for only 14.5% of the global population, it carries ~30% of the global disease burden (Alessandrini and Pepper, 2014). Sub-Saharan Africa (SSA) in particular carries over 90% of the world's malaria burden and related deaths (World Health Organization, 2020a), as well as ~25% of the world's Tuberculosis (TB) burden (World Health Organization, 2020b). Furthermore, ~70% of individuals infected with Human immunodeficiency virus (HIV) are located in SSA (Avert, 2020). In addition, SSA is experiencing a rapidly increasing epidemic of non-communicable diseases (NCDs) including cardiovascular diseases, diabetes mellitus, respiratory diseases, cancer and neuropsychiatric disorders (Mayosi *et al.*, 2009; Nyirenda, 2016). Even though the burden of infectious disease currently outweighs that of NCDs in SSA, the burden of

NCDs is expected to overtake that of infectious diseases by 2035 (Nyirenda, 2016). The following sections will detail the drugs frequently prescribed for common diseases in SSA, as well as the ADME genes involved in their metabolism.

1.5.1 Antiretrovirals (ARVs)

Various ARV drugs are currently available in Africa, including but not limited to efavirenz, dolutegravir, raltegravir, nevirapine, ritonavir and lopinavir (Meintjes *et al.*, 2017). Additionally, ARVs are most commonly administered together in a single pill called a Fixed Dose Combination (FDC) with the aim of reducing the number of pills an individual needs to take, thereby increasing adherence to treatment (Meintjes *et al.*, 2017; Clinical Info HIV.gov, 2021). Two of the most common FDCs for first-line treatment include tenofovir and emtricitabine in combination with either efavirenz or, more recently, dolutegravir (Meintjes *et al.*, 2017).

The World Health Organization (WHO) guidelines previously recommended efavirenz-based first-line Antiretroviral Treatment (ART) (World Health Organization, 2018), and as such, this drug is available in public sector programs in most countries in southern Africa (Meintjes *et al.*, 2017). Efavirenz is metabolised primarily by CYP2B6, while CYP3A4, CYP3A5, CYP2C9, CYP1A1, CYP2A6 and UGT2B7 provide a minor contribution to its metabolism (*Table 1.2*) (Bains, 2013; Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017). Subsequently, dolutegravir has been shown to be superior to efavirenz-based treatment due to its high potency and barriers to resistance, as well as good tolerability (Phillips *et al.*, 2019). For these reasons the WHO updated its guidelines, recommending the use of dolutegravir in first-line treatment regimens (World Health Organization, 2018; Phillips *et al.*, 2019). However, this drug is primarily available in the private sector, with only a few countries in SSA offering it in the public sector (Meintjes *et al.*, 2017). Nevertheless, it is likely to become more accessible in the public sector in the future, when more affordable generic options are available on a larger scale (Meintjes *et al.*, 2017). In some treatments, dolutegravir is replaced by raltegravir (Meintjes *et al.*, 2017), both of which are metabolised primarily by UGT1A1 (*Table 1.2)* (Kassahun *et al.*, 2007; Castellino *et al.*, 2013; Elliot, Mahungu and Owen, 2017).

Table 1.2: Major and minor genes involved in the ADME of commonly used antiretrovirals, antimalarial and anti-TB drugs.

Drug Class	Drug Name	Major ADME genes involved	Minor ADME genes involved	References
Antiretroviral	Efavirenz	*CYP2B6*	*CYP3A4, CYP3A5, CYP2C9, CYP1A1, CYP2A6, UGT2B7*	(Bains, 2013; Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017)
	Nevirapine	*CYP2B6, CYP3A4*	*CYP3A5, GSTM1*	(Bains, 2013; Dandara *et al.*, 2014; Ciccacci *et al.*, 2017; Elliot, Mahungu and Owen, 2017)
	Dolutegravir	*UGT1A1*		(Castellino *et al.*, 2013; Elliot, Mahungu and Owen, 2017)
	Raltegravir	*UGT1A1*		(Kassahun *et al.*, 2007; Elliot, Mahungu and Owen, 2017)
	Ritonavir	*CYP3A4, CYP3A5*		(Bains, 2013; Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017)
	Lopinavir	*CYP3A4, CYP3A5*	*SLOCO1B1*	(Bains, 2013; Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017)
Antimalarial	Chloroquine	*CYP2C9, CYP3A4, CYP3A5*	*CYP1A2, CYP2A6, CYP2B6, CYP2C19, CYP2D6*	(Bains, 2013; Dandara *et al.*, 2014)
	Amodiaquine	*CYP2C8*	*CYP2D6, CYP3A4*	(Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017)
	Mefloquine	*CYP3A4, CYP1A2*	*ABCB1, CYP2A6, CYP2B6, CYP2C9, CYP2C19*	Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017)
	Quinine	*CYP3A4, CYP3A5*	*CYP2B6*	(Bains, 2013; Elliot, Mahungu and Owen, 2017; Zhou *et al.*, 2019)
	Artemether	*CYP3A4, CYP3A5, CYP2A6, UGT1A9*	*CPY2B6, UGT2B7*	(Navaratnam *et al.*, 2000; Ilett *et al.*, 2002; Ali *et al.*, 2010; Bains, 2013; Dandara *et al.*, 2014; Zang *et al.*, 2014; Elliot, Mahungu and Owen, 2017)
	Lumefantrine	*CYP3A4, CYP3A5*	*CYP2B6, CYP1A2, CYP2C9, CYP2C19*	(Dandara *et al.*, 2014; Abdullahi *et al.*, 2020)
	Artesunate	*CYP3A4, CYP3A5, CYP2A6, UGT1A9*	*CPY2B6, UGT2B7*	(Ilett *et al.*, 2002; Zang *et al.*, 2014; Elliot, Mahungu and Owen, 2017)
Anti-TB	Isoniazid	*NAT2, CYP2E1, GSTM1, GSTT1*		(Bains, 2013; Wang *et al.*, 2016; Elliot, Mahungu and Owen, 2017)
	Rifampicin	*CYP3A4, CYP3A5, CYP2B6*	*CYP2C8, CYP2C9, UGTs, GSTs*	(Elliot, Mahungu and Owen, 2017; Sloan *et al.*, 2017)

Another option in first-line HIV treatment is nevirapine; however, this is no longer recommended for new patients starting ART due to the severe toxicity that is often associated with its use (Meintjes *et al.*, 2017). Nevirapine is primarily metabolised by CYP2B6 and CYP3A4, with contributions from CYP3A5 and GSTM1 (*Table 1.2)* (Bains, 2013; Dandara *et al.*, 2014; Ciccacci *et al.*, 2017; Elliot, Mahungu and Owen, 2017).

When first-line treatment fails due to drug resistance, second-line ARV treatment is prescribed (Meintjes *et al.*, 2017). ARV drugs prescribed as second-line treatment include ritonavir and lopinavir (Meintjes *et al.*, 2017). Both drugs are metabolised by CYP3A4 and CYP3A5 (Ripudaman K Bains, 2013; Dandara *et al.*, 2014), and recently, SLCO1B1 has been associated with lopinavir plasma concentration (*Table 1.2)* (Elliot, Mahungu and Owen, 2017).

Each of these drugs are associated with some common ADRs, ranging from rash and gastrointestinal problems, to hepatitis and central nervous system symptoms (Meintjes *et al.*, 2017). These ADRs may be in part due to genetic variation within the genes that metabolise the commonly prescribed ARVs. Additionally, most ARV compounds are transported by ABCB1 and ABCC2 proteins, and variations within transporter genes can influence the safety and efficacy of the drugs they transport (Kis *et al.*, 2016). Therefore, due to the large number of individuals receiving ARV treatment in SSA, CNVs within these genes are an important consideration (Williamson *et al.*, 2013).

1.5.2 Anti-malarial drugs

Malaria remains an overwhelming issue in SSA, which carries over 90% of the global malaria burden, as well as 94% of all malaria-related deaths (Conrad and Rosenthal, 2019; World Health Organization, 2019). Large efforts have been made in order to prevent infection as well as provide treatment, and, in 2018, 214 million malaria treatment courses were delivered, 98% of which were in the WHO Africa region (World Health Organization, 2019). Traditional malaria treatments included chloroquine, quinine, sulphadoxine-pyrimethamine, amodiaquine and mefloquine which were the most frequently prescribed first-line therapies for *Plasmodium falciparum* (*P.f*) up until 2000, and even later until the end of the 2000s in some SSA counties (Flegg *et al.*, 2013). As resistance to these monotherapies emerged and spread, resulting in increased malaria mortality, the WHO changed its guidelines and recommend that artemisinin-based combination therapies (ACTs) be used for the treatment of uncomplicated *P.f.* malaria (Ogutu, 2013). The use of a combination of different anti-malarial drugs, with different modes of action, was thought to hinder the emergence of resistance to the drugs (Ogutu, 2013). In SSA, the most common ACTs that are prescribed are (1) artemether in combination with lumefantrine (AL), with 30 of the 47 SSA countries using it as first-line therapy, and 8 as second-line therapy;

(2) artesunate in combination with amodiaquine (ASAQ); and (3) artesunate in combination with mefloquine (ASMQ) (Ogutu, 2013; Elliot, Mahungu and Owen, 2017; Conrad and Rosenthal, 2019).

Artemether and artesunate are primarily metabolised by CYP3A4, CYP3A5 and CYP2A6, with minor contributions from CYP2B6, into dihydroartemisinin (*Table 1.2)* (Navaratnam *et al.*, 2000; Ali *et al.*, 2010; Bains, 2013; Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017). Dihydroartemisinin is in turn metabolised by UGT1A9, with a minor role played by UGT2B7 (*Table 1.2)* (Ilett *et al.*, 2002; Zang *et al.*, 2014). Lumefantrine is primarily metabolised by CYP3A4 and CYP3A5, with contributions from CYP2B6, CYP1A2, CYP2C9 and CYP2C19 (*Table 1.2)* (Dandara *et al.*, 2014; Abdullahi *et al.*, 2020). Amodiaquine is primarily metabolised by CYP2C8, with contributions from CYP2D6 and CYP3A4, while mefloquine is primarily metabolised by CYP3A4 and CYP1A2, with contributions from ABCB1, CYP2A6, CYP2B6, CYP2C9 and CYP2C19 (*Table 1.2)* (Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017).

Although ACT is the recommended treatment for uncomplicated *P.f.* malaria infection, ineffective monotherapies such as quinine and chloroquine continue to be used for treatment, predominantly due to the high cost and frequent lack of availability of ACT (Wasunna *et al.*, 2008; Achan *et al.*, 2011; Adedeji *et al.*, 2011). Additionally, while artesunate monotherapy is recommended for severe *P.f*. malaria (Li and Weina, 2010), quinine is the most common drug used for treatment (Li and Weina, 2010). Both quinine and chloroquine are metabolised predominantly by CYP3A4 and CYP3A5 (Bains, 2013; Elliot, Mahungu and Owen, 2017; Zhou *et al.*, 2019), with CYP2C9 as well as CYP1A2, CYP2A6, CYP2B6, CYP2C19 and CYP2D6 also contributing to the metabolism of chloroquine (*Table 1.2)* (Bains, 2013; Dandara *et al.*, 2014).

Antimalarial drugs exhibit a wide range of inter-individual variability and commonly cause ADRs (Gebreyohannes *et al.*, 2017; Ndagije *et al.*, 2018), which makes genetic variation within the above mentioned genes an important consideration.

1.5.3 Anti-TB drugs

SSA carries one quarter of the global burden of TB (World Health Organization, 2020b), and TB is the most common opportunistic infection and cause of death among

HIV-infected individuals (Manosuthi, Wiboonchutikul and Sungkanuparph, 2016). The currently recommended standard treatment for drug-susceptible TB includes a combination of the following four drugs: isoniazid, rifampicin, pyrazinamide and ethambutol (Brito *et al.*, 2014; Manosuthi, Wiboonchutikul and Sungkanuparph, 2016). NAT2, CYP2E1 as well as GSTM1 and GSTT1, are involved in the metabolism of Isoniazid (*Table 1.2)* (Bains, 2013; Brito *et al.*, 2014; Wang *et al.*, 2016; Elliot, Mahungu and Owen, 2017). Rifampicin is metabolized by enzymes of the CYP450 system (Arbex *et al.*, 2010), and is also a potent inducer of various ADME genes including, but not limited to, CYP3A4, CYP3A5, CYP2B6, CYP2C8 and CYP2C9 (Bibi, 2008; Bains, 2013; Cho *et al.*, 2016), as well as UGTs and GSTs (*Table 1.2).* When rifampicin induces the enzyme activity of the CYP450 enzymes, it increases its own metabolism (Bibi, 2008). Ethambutol is metabolised by ADHs and ALDHs which are Phase I metabolisers (Sarkar and Ganguly, 2016). Lastly, pyrazinamide is metabolised into its active form by hepatic microsomal deamidase and is thus not relevant to this study (Sarkar and Ganguly, 2016; Chirehwa *et al.*, 2017).

The above mentioned drugs display a large amount of inter-individual variability and are associated with various ADRs including hepatotoxicity (Brito *et al.*, 2014; Sarkar and Ganguly, 2016). Genetic variation within these genes may play a role in the variable response and ADRs associated with these drugs (Sarkar and Ganguly, 2016).

1.5.4 Anticoagulants

Medical conditions that require the use of anticoagulant treatment are important causes of morbidity and mortality worldwide (Semakula *et al.*, 2021). Reduced childhood mortality as well as an improvement in the control of infectious diseases such as HIV have led to an increase in life expectancy in SSA (Semakula *et al.*, 2021). This increased life expectancy is associated with various cardiovascular diseases and consequently an increased requirement for the use of anticoagulant drugs in the region (Semakula *et al.*, 2021). Warfarin is an anticoagulant used for the treatment and prophylaxis of various types of blood clots (Semakula *et al.*, 2020). It is administered as a combination of (R)-warfarin and (S)-warfarin, with (S)-warfarin being two to five times more potent that (R)-warfarin (Al-Eitan, Almasri and Khasawneh, 2018). CYP2C9 is considered to be the most important enzyme in warfarin metabolism as it is the primary pathway for the elimination of (S)-warfarin. Individuals with reduced CYP2C9 enzyme activity are classified at PMs and are at a higher risk of serious side

effects, including fatal bleeding (Takahashi and Echizen, 2003). Therefore, PMs often require a smaller dose to avoid toxicity, as the drug is metabolised at a slower rate (Takahashi and Echizen, 2003). Warfarin is an inexpensive drug and is therefore commonly used in public health care systems (Laäs and Naidoo, 2018). As it is most commonly prescribed oral anticoagulant in SSA, deletions of or within this gene are important to consider (Semakula *et al.*, 2020).

1.5.5 Lipid-lowering drugs

The use of lipid-lowering drugs, also known as statins, on the African continent is expected to increase as the burden of cardiovascular disease increases. This increase in cardiovascular diseases is primarily due to a change in lifestyles across the continent as well an increased burden of HIV-associated dyslipidaemia linked to ARV treatment (Soko, Masimirembwa and Dandara, 2018). The use of statins has been shown to decrease the morbidity and mortality associated with cardiovascular diseases by reducing the total circulating levels of cholesterol and low-density lipoprotein in the blood (Soko, Masimirembwa and Dandara, 2018; Yan *et al.*, 2018). Rosuvastatin and fluvastatin are known as CYP2C9-metabolized statins (Scripture and Pieper, 2001; Olsson, McTaggart and Raza, 2002; Yan *et al.*, 2018), while atorvastatin is more extensively metabolized by CYP3A4 (Korhonova, Doricakova and Dvorak, 2015; Yan *et al.*, 2018). Genetic variation within these genes, including CNVs, may have an effect on the drugs' lipid-lowering efficacy.

1.5.6 Anti-cancer drugs

Cancer is an increasing global public health burden, particularly in SSA where there are an estimated 1 million new cases of cancer every year and this is expected to double by 2030 (Dent *et al.*, 2017). Breast and cervical cancer in females and prostate, liver and colorectal cancer in males are the most common cancer types in SSA (Bahnassy, Abdellateif and Zekri, 2020). However, the majority of chemotherapy drugs are not available in low- to middle-income countries (LMICs), due to their high costs as well as difficulties associated with administering and monitoring them in low-resource settings (Boyle *et al.*, 2019). In mid-2020, various pharmaceutical companies agreed to expand access to 20 chemotherapy drugs in 26 SSA countries through cost reductions of ~60% (Clinton Health Access Initiative, 2020). The drugs included in the agreement cover the recommended regimes for 27 cancer types and allow for

complete therapy of breast, cervical and prostate cancer, the three cancer types that cause the most deaths in Africa (The Cancer Atlas, 2018; Clinton Health Access Initiative, 2020). Some of the chemotherapy agents included in the agreement include cisplatin, docetaxel, fluorouracil, oxaliplatin, paclitaxel and tamoxifen (Cancer, 2021). The metabolism of tamoxifen involves multiple enzymes, including CYP2D6, CYP3A4, CYP3A5, CYP2C9 and CYP2C19 (Cronin-Fenton, Damkier and Lash, 2014). Furthermore SULT1A1 and UGT enzymes, UGT1A8, UGT1A10 and UGT2B7, are responsible for the conversion of tamoxifen metabolites into excretable forms (Cronin-Fenton, Damkier and Lash, 2014). Paclitaxel is prominently metabolised by CYP2C8 and CYP3A4 enzymes (Spratlin and Sawyer, 2007), while docetaxel is metabolised by CYP3A4 and CYP3A5 enzymes (Kenmotsu and Tanigawara, 2015). The metabolism of platinum-based drugs, cisplatin and oxaliplatin, is not primarily mediated by enzymes (Lentz *et al*., 2005; Allocati *et al.*, 2018); however, GST enzymes, GSTM1 and GSTT1, can catalyse one of the reactions (Allocati *et al.*, 2018). Lastly, the metabolism of fluorouracil is metabolised by DPD, which is expressed by the *DPYD* gene (Lentz *et al*., 2005).

1.6 Pharmacogenomics in SSA

Due to the high prevalence of disease, the public health systems in SSA countries are under enormous pressure to provide treatment options that are as efficient and cost effective as possible (Warnich *et al.*, 2011). However, available treatments are often associated with toxicity and ADRs, which are responsible for a large proportion of hospitalisations, with up to 6.3% of hospital admissions in SSA being a direct consequences of an ADR and up to 49.5% of hospitalized patients in SSA developing an ADR during their stay (Wester *et al.*, 2008; Appiah, 2012; Sim, Kacevska and Ingelman-Sundberg, 2013). Not only do ADRs place a significant burden on the public health system, but they also play a role in patient non-compliance, which further exacerbates the high burden of disease in the region (Warnich *et al.*, 2011).

It is well known that African populations display a large amount of diversity owing to the age of the population, and vast genetic differences are observed both between different African populations as well as between African populations and those outside of Africa (Alessandrini and Pepper, 2014). Therefore, drugs that have been developed based on European populations may not be effective or may have toxic effects in

African populations. The fact that the majority of drugs are developed based on European populations and are administered to African patients without further testing regarding efficacy and toxicity (Matimba, Dhoro and Dandara, 2016) may provide an explanation for the high ADR rates observed in African populations.

The investigation of genetic variation within pharmacogenes in African populations could provide more information regarding drug efficacy, safety as well as relevant dosages for African patients. For example, the recent study performed by Santos and colleagues found that in African individuals, deletions and duplications in *CYP2B6* were the most frequent (Santos *et al.*, 2018). *CYP2B6* is known to be involved in the ADME of many drugs, including antiretroviral drugs efivarenz and nevaripine, artemether, used for the treatment of malaria, as well as rifampicin, which is used for the treatment of TB (Bains, 2013). Therefore, African-specific polymorphisms within *CYP2B6* that may alter its ability to metabolise these drugs are extremely important to consider due to the high burden of these diseases in the region (Zanger and Klein, 2013). This is highlighted by the recommendation that HIV patients in SSA require a reduced dose of efavirenz due to genetic variation within the *CYP2B6* gene (Mukonzo *et al.*, 2014; Dhoro *et al.*, 2015).

Furthermore, pharmacogenetic markers identified in European populations may not be suitable in an African setting. For example, the AmpliChip CYP450 was the first molecular diagnostic test to be approved by the US Food and Drug Administration (FDA) with the aim of practicing pharmacogenetic testing in a clinical setting (He, Hoskins and McLeod, 2011). The test is used for the analysis of SNPs and CNVs found in the *CYP2D6* and *CYP2C19* genes that have been shown to be involved in the ADME of drugs used in the treatment of depression, gastroesophageal reflux and cardiovascular diseases (He, Hoskins and McLeod, 2011). However, it must be noted that the AmpliChip CYP450 has not been validated in any African population living in Africa (Dodgen *et al.*, 2013). When the AmpliChip CYP450 was assessed in a group of South African individuals, the success rate was low and a large number of unknown predicted phenotype calls were observed, indicating the limited utility of this test in an African setting (Dodgen *et al.*, 2013). This highlights the importance of identifying pharmacogenetic markers specific to African populations to ensure the success of clinical pharmacogenetics in an African setting.

A study carried out by Mitchell and colleagues assessed genetic variants within the *CYP2C9* and *VKORC1* genes, which are known to play a role in the variable response to Warfarin (Mitchell, Gregersen and Krause, 2011). Thirty four variants had previously been identified within the *CYP2C9* gene based on “Caucasian” (i.e. European, American or similar), African-American and “Asian” (i.e. Central Asian, East Asian, South Asian, Southeast Asian or Western Asian) populations (Mitchell, Gregersen and Krause, 2011). Results from this study showed that only seven of these 34 variants were observed in black South African individuals and an additional 26 novel *CYP2C9* SNPs were observed (Mitchell, Gregersen and Krause, 2011). This further illustrates the limited utility of pharmacogenetic markers identified in non-African populations being used in an African setting.

There is a dire need to address the growing burden of disease in Africa, both in terms of infectious disease such as HIV and TB, as well as non-communicable diseases such as cardiovascular disease, diabetes and cancer. In order to address these issues, we require an improved understanding of various aspects including the molecular mechanisms of disease susceptibility and response to drugs. The vast genetic diversity present in African populations provides a clear benefit in that it could allow for the improved understanding of individual variability in disease and drug response (Matimba, Dhoro and Dandara, 2016). With the majority of pharmacogenomic studies being performed on “Caucasian” (i.e. European, American or similar) patients, only a small percentage of overall human genetic diversity is being assessed (Matimba, Dhoro and Dandara, 2016). Therefore, pharmacogenomic studies based in Africa would analyse populations with the greatest genetic diversity, thereby increasing the likelihood of identifying relevant, as well as novel, pharmacogenomic variation (Matimba, Dhoro and Dandara, 2016).

1.7 Rationale

Widespread pharmacogenetic profiling of African populations will aid in significantly improving patient care on the continent (Alessandrini and Pepper, 2014). Populations in SSA would benefit greatly from the optimisation of drug therapies through the use of pharmacogenetics, which would increase the efficacy of treatment as well as reduce ADRs, thereby decreasing the overall cost of treatment and disease burden in the region (Warnich *et al*., 2011). While pharmacogenetically relevant SNVs in ADME genes have been well characterised, CNV is an important, yet understudied source of variation, that needs to be further investigated in order to obtain a more complete set of pharmacogenetically relevant variants in the African context.

1.8 Study Aim and Objectives

1.8.1 Aim

Due to a paucity of studies focused on CNV in ADME genes in SSA populations where the burden of disease and adverse drug reactions are high, the aim of this study is to perform a systematic assessment of the CNV landscape in ADME genes in SSA populations.

1.8.2 Objectives

1. To call CNVs from short-read whole genome sequence data using a variety of CNV calling tools including GenomeSTRiP, Manta, Delly, Lumpy and GATK.
2. To validate CNVs called from short-read sequence data using PacBio long-read sequencing.
3. To assess whether short-read sequencing is adequate for accurate CNV calls.
4. To determine which CNVs can be called from the Infinium H3Africa Consortium Array v2, thus allowing for the assessment of the utility of the array to call CNV's.
5. To select a list of ADME CNVs that are the most relevant in an African setting

1.9 Materials and Methods

The materials and methods used will be discussed in each of the results chapters.

1.10 Thesis outline

The format of this thesis is the ‘Divided Block’ format, whereby Chapter 2 focusses on a short-read sequencing approach for CNV calling, Chapter 3 focusses on a long-read sequencing approach for CNV calling and Chapter 4 focusses on an array-based approach for CNV calling. Each chapter contains its own ‘Introduction’, ‘Methods and Materials’, ‘Results’ and ‘Discussion’ sections. Although the chapters are distinct, some of the results are compared across chapters. Chapter 5 is the final conclusion chapter which integrates the data from the individual results chapters and discusses the overall conclusions that emerge from this thesis, as well as future work that can be done. All of the custom scripts used in this study as well as supplementary files can be found in the following GitHub repository: https://github.com/lauracottino/PhD_Thesis.

1.11 Reference list

Abdullahi, S.T., Soyinka, J.O., Olagunju, A., *et al.* (2020) CYP2B6*6 Genotype Specific Differences in Artemether-Lumefantrine Disposition in Healthy Volunteers. *The Journal of Clinical Pharmacology*. 60(3), pp. 351–360. doi:https://doi.org/10.1002/jcph.1527.

Abyzov, A., Urban, A.E., Snyder, M., *et al.* (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*. 21(6), pp. 974–984. doi:10.1101/gr.114876.110.

Achan, J., Talisuna, A.O., Erhart, A., *et al.* (2011) Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malaria Journal*. 10, pp. 144. doi:10.1186/1475-2875-10-144.

Adedeji, A., Sanusi, B., Tella, A., *et al.* (2011) Exposure to anti-malarial drugs and monitoring of adverse drug reactions using toll-free mobile phone calls in private retail sector in Sagamu, Nigeria: Implications for pharmacovigilance. *Malaria Journal*. 10, pp. 230. doi:10.1186/1475-2875-10-230.

Ahmed, S., Zhou, Z., Zhou, J., *et al.* (2016) Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine. *Genomics, Proteomics and Bioinformatics*. 14(5), pp. 298–313. doi:10.1016/j.gpb.2016.03.008.

Aka, I., Bernal, C.J., Carroll, R., *et al.* (2017) Clinical Pharmacogenetics of Cytochrome P450-Associated Drugs in Children. *Journal of Personalized Medicine*. 7(4), pp. 14. doi:10.3390/jpm7040014.

Akhondzadeh, S. (2014) Personalized medicine: a tailor made medicine. *Avicenna Journal of Medical Biotechnology*. 6(4), pp. 191.

Al-Eitan, L.N., Almasri, A.Y. and Khasawneh, R.H. (2018) Impact of CYP2C9 and VKORC1 Polymorphisms on Warfarin Sensitivity and Responsiveness in Jordanian Cardiovascular Patients during the Initiation Therapy. *Genes*. 9(12), pp. 578. doi:10.3390/genes9120578.

Alessandrini, M. and Pepper, M.S. (2014) Priority pharmacogenetics for the African continent: focus on CYP450. *Pharmacogenomics*. 15(3), pp. 385–400. doi:10.2217/pgs.13.252.

Ali, S., Najmi, M.H., Tarning, J., *et al.* (2010) Pharmacokinetics of artemether and dihydroartemisinin in healthy Pakistani male volunteers treated with artemether-lumefantrine. *Malaria Journal*. 9, pp. 275. doi:10.1186/1475-2875-9-275.

Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping. *Nature Reviews Genetics*. 12(5), pp. 363–376. doi:10.1038/nrg2958.

Allied Against Cancer (2021) *Cancer Access Partnership*. Available at: https://www.alliedagainstcancer.org/access-partnership (Accessed: 2 May 2021).

Allocati, N., Masulli, M., Di Ilio, C., *et al.* (2018) Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases. *Oncogenesis*. 7(1), pp. 8. doi:10.1038/s41389-017-0025-3.

Amarasinghe, S.L., Su, S., Dong, X., *et al.* (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*. 21(1), pp. 30. doi:10.1186/s13059-020-1935-5.

Appiah, B. (2012) Africa struggles to improve drug safety. *Canadian Medical Association Journal*. 184(10), pp. 533–534. doi:10.1503/cmaj.109-4199.

Arbex, M.A., Varella, M. C.L., de Siqueira, H.R., *et al.* (2010) Antituberculosis drugs: drug interactions, adverse effects, and use in special situations. Part 1: first-line drugs. *Jornal Brasileiro de Pneumologia*. 36, pp. 626–640. doi:10.1590/s1806-

37132010000500016.
Ardui, S., Ameur, A., Vermeesch, J.R., *et al.* (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*. 46(5), pp. 2159–2168. doi:10.1093/nar/gky066.
Arlt, M.F., Wilson, T.E. and Glover, T.W. (2012) Replication stress and mechanisms of CNV formation. *Current Opinion in Genetics & Development*. 22(3), pp. 204–210. doi:10.1016/j.gde.2012.01.009.
Avert (2020) *HIV and AIDS in East and Southern Africa Regional Overview*. Available at: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/overview (Accessed: 3 November 2019).
Bahnassy, A.A., Abdellateif, M.S. and Zekri, A.N. (2020) Cancer in Africa: Is It a Genetic or Environmental Health Problem? *Frontiers in Oncology*. 10, pp. 2508. doi:10.3389/fonc.2020.604214.
Bains, R.K. (2013) African variation at Cytochrome P450 genes: Evolutionary aspects and the implications for the treatment of infectious diseases. *Evolution, Medicine, and Public Health*. 2013(1), pp. 118–134. doi:10.1093/emph/eot010.
Bibi, Z. (2008) Role of cytochrome P450 in drug interactions. *Nutrition & Metabolism*. 5(1), pp. 27. doi:10.1186/1743-7075-5-27.
Boyle, P., Ngoma, T., Sullivan, R., *et al.* (2019) Cancer in Africa: the way forward. *Ecancer Medical Science*. 13, pp. 953. doi:10.3332/ecancer.2019.953.
Bradford, L.D. (2002) CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants. *Pharmacogenomics*. 3(2), pp. 229–243. doi:10.1517/14622416.3.2.229.
Brito, T.C., Possuelo, L.G., Valim, A.R.M., *et al.* (2014) Polymorphisms in CYP2E1, GSTM1 and GSTT1 and anti-tuberculosis drug-induced hepatotoxicity. *Anais da Academia Brasileira de Ciências*. 86, pp. 855–865. doi: 10.1590/0001-3765201420130350.
Buysse, K., Delle Chiaie, B., Van Coster, R., *et al.* (2009) Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *European Journal of Medical Genetics*. 52(6), pp. 398–403. doi:10.1016/j.ejmg.2009.09.002.
Cabianca, D.S. and Gabellini, D. (2010) FSHD: copy number variations on the theme of muscular dystrophy. *The Journal of Cell Biology*. 191(6), pp. 1049–1060. doi:10.1083/jcb.201007028.
Carter, N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*. 39(7), pp. 16-21. doi:10.1038/ng2028.
Castellino, S., Moss, L., Wagner, D., *et al.* (2013) Metabolism, excretion, and mass balance of the HIV-1 integrase inhibitor dolutegravir in humans. *Antimicrobial Agents and Chemotherapy*. 57(8), pp. 3536–3546. doi:10.1128/AAC.00292-13.
Caudle, K.E., Dunnenberger, H.M., Freimuth, R.R., *et al.* (2017) Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genetics in Medicine*. 19(2), pp. 215–223. doi:10.1038/gim.2016.87.
Chaisson, M.J.P., Wilson, R.K. and Eichler, E.E. (2015) Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*. 16(11), pp. 627–640. doi:10.1038/nrg3933.
Chen, K., Wallis, J.W., McLellan, M.D., *et al.* (2009) BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*. 6(9), pp. 677–681. doi:10.1038/nmeth.1363.
Chen, X., Schulz-Trieglaff, O., Shaw, R., *et al.* (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 32(8), pp. 1220–1222. doi:10.1093/bioinformatics/btv710.

Chirehwa, M.T., McIlleron, H., Rustomjee, R., *et al.* (2017) Pharmacokinetics of Pyrazinamide and Optimal Dosing Regimens for Drug-Sensitive and -Resistant Tuberculosis. *Antimicrobial Agents and Chemotherapy*. 61(8), pp. 2171–2179. doi:10.1128/AAC.00490-17.

Cho, D.Y., Shen, J.H.Q., Lemler, S.M., *et al.* (2016) Rifampin enhances cytochrome P450 (CYP) 2B6-mediated efavirenz 8-hydroxylation in healthy volunteers. *Drug Metabolism and Pharmacokinetics*. 31(2), pp. 107–116. doi:10.1016/j.dmpk.2015.07.002.

Ciccacci, C., Latini, A., Politi, C., *et al.* (2017) Impact of glutathione transferases genes polymorphisms in nevirapine adverse reactions: a possible role for GSTM1 in SJS/TEN susceptibility. *European Journal of Clinical Pharmacology*. 73(10), pp. 1253–1259. doi:10.1007/s00228-017-2295-2.

Clarke, J., Wu, H.C., Jayasinghe, L., *et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*. 4(4), pp. 265–270. doi:10.1038/nnano.2009.12.

Clinton Health Access Initiative (2020) *New agreements to lower prices and increase access to lifesaving cancer treatment in sub-Saharan Africa*. Available at: https://www.clintonhealthaccess.org/agreements-to-increase-access-to-cancer-treatment-in-africa/ (Accessed: 2 May 2021).

Colella, S., Yau, C., Taylor, J.M., *et al.* (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*. 35(6), pp. 2013–2025. doi:10.1093/nar/gkm076.

Conrad, D.F., Andrews, T.D., Carter, N.P., *et al.* (2006) A high-resolution survey of deletion polymorphism in the human genome. *Natur Genetics*. 38(1), pp. 75–81. doi:10.1038/ng1697.

Conrad, M.D. and Rosenthal, P.J. (2019) Antimalarial drug resistance in Africa: the calm before the storm? *The Lancet Infectious Diseases*. 19(10), pp. 338–351. doi:10.1016/S1473-3099(19)30261-0.

Crews, K.R., Hicks, J.K., Pui, C.H., *et al.* (2012) Pharmacogenomics and individualized medicine: translating science into practice. *Clinical Pharmacology and Therapeutics*. 2012/09/05. 92(4), pp. 467–475. doi:10.1038/clpt.2012.120.

Cronin-Fenton, D.P., Damkier, P. and Lash, T.L. (2014) Metabolism and transport of tamoxifen in relation to its effectiveness: new perspectives on an ongoing controversy. *Future Oncology*. 10(1), pp. 107–122. doi:10.2217/fon.13.168.

Dandara, C., Swart, M., Mpeta, B., *et al.* (2014) Cytochrome P450 pharmacogenetics in African populations: implications for public health. *Expert Opinion on Drug Metabolism & Toxicology*. 10(6), pp. 769–785. doi:10.1517/17425255.2014.894020.

Delaneau, O., Howie, B., Cox, A.J., *et al.* (2013) Haplotype estimation using sequencing reads. *American Journal of Human Genetics*. 93(4), pp. 687–696. doi:10.1016/j.ajhg.2013.09.002.

Dent, J., Manner, C.K., Milner, D., *et al.* (2017) *Africa's Emerging Cancer Crisis: A Call to Action*. Available at: https://bvgh.org/wp-content/uploads/2017/07/Africas-Emerging-Cancer-Crisis-A-Call-to-Action.pdf (Accessed: 2 May 2021).

Dhoro, M., Zvada, S., Ngara, B., *et al.* (2015) CYP2B6*6, CYP2B6*18, Body weight and sex are predictors of efavirenz pharmacokinetics and treatment response: population pharmacokinetic modeling in an HIV/AIDS and TB cohort in Zimbabwe. *BMC Pharmacology & Toxicology*. 16, pp. 4. doi:10.1186/s40360-015-0004-2.

Dodgen, T.M., Hochfeld, W.E., Fickl, H., *et al.* (2013) Introduction of the AmpliChip CYP450 Test to a South African cohort: A platform comparative prospective cohort study. *BMC Medical Genetics*. 14(1). pp. 1–15. doi:10.1186/1471-2350-14-20.

Eichler, E.E. (2019) Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *The New England Journal of Medicine*. 381(1), pp. 64–74. doi:10.1056/NEJMra1809315.

Eid, J., Fehr, A., Gray, J., *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*. 323(5910), pp. 133–138. doi:10.1126/science.1162986.

Elliot, E., Mahungu, T. and Owen, A. (2017) 'Current Progress in the Pharmacogenetics of Infectious Disease Therapy' in Tibayrenc, M. (ed.) *Genetics and Evolution of Infectious Disease*. London: Elsevier. pp. 435–457. doi:https://doi.org/10.1016/B978-0-12-799942-5.00019-6.

English, A.C., Salerno, W.J. and Reid, J.G. (2014) PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*. 15, pp. 180. doi:10.1186/1471-2105-15-180.

Feuk, L., Carson, A.R. and Scherer, S.W. (2006) Structural variation in the human genome. *Nature Reviews Genetics*. 7(2), pp. 85–97. doi:10.1038/nrg1767.

Fiegler, H., Redon, R., Andrews, D., *et al.* (2006) Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Research*. 2006/11/22. 16(12), pp. 1566–1574. doi:10.1101/gr.5630906.

Flegg, J.A., Metcalf, C.J.E., Gharbi, M., *et al.* (2013) Trends in antimalarial drug use in Africa. *The American Journal of Tropical Medicine and Hygiene*. 89(5), pp. 857–865. doi:10.4269/ajtmh.13-0129.

Freeman, J.L., Perry, G.H., Feuk, L., *et al.* (2006) Copy number variation: New insights in genome diversity. *Genome Research*. 16(8), pp. 949–961. doi:10.1101/gr.3677206.

Fuselli, S., de Filippo, C., Mona, S., *et al.* (2010) Evolution of detoxifying systems: the role of environment and population history in shaping genetic diversity at human CYP2D6 locus. *Pharmacogenetics and Genomics*. 20(8), pp. 485–499. doi:10.1097/FPC.0b013e32833bba25.

Gaedigk, A. and Leeder, J.S. (2012) CYP2D6, SULT1A1 and UGT2B17 copy number variation: quantitative detection by multiplex PCR. *Pharmacogenomics.* 13, pp. 91–111. doi: 10.2217/pgs.11.135.

Gallagher, C.J., Kadlubar, F.F., Muscat, J.E., *et al.* (2007) The UGT2B17 gene deletion polymorphism and risk of prostate cancer. A case-control study in Caucasians. *Cancer Detection and Prevention*. 31(4), pp. 310–315. doi:10.1016/j.cdp.2007.07.005.

Garte, S., Gaspari, L., Alexandrie, A.K., *et al.* (2001) Metabolic gene polymorphism frequencies in control populations. *Cancer Epidemiology, Biomarkers & Prevention*. 10(12), pp. 1239–1248.

Gebreyohannes, E.A., Bhagavathula, A.S., Seid, M.A., *et al.* (2017) Antimalarial Treatment Outcomes In Ethiopia: A Systematic Review And Meta-Analysis. *Value in Health*. 20(9), pp. 795. doi:10.1016/j.jval.2017.08.2346.

Genovese, G., Handsaker, R.E., Li, H., *et al.* (2013) Using population admixture to help complete maps of the human genome. *Nature Genetics*. 45(4), pp. 406–414. doi:10.1038/ng.2565.

Goh, L.L., Lim, C.W., Sim, W.C., *et al.* (2017) Analysis of Genetic Variation in CYP450 Genes for Clinical Implementation. *PLoS One*. 12(1), pp. 1–16. doi:10.1371/journal.pone.0169233.

Griese, E.U., Asante-Poku, S., Ofori-Adjei, D., *et al.* (1999) Analysis of the CYP2D6 gene mutations and their consequences for enzyme function in a West African population. *Pharmacogenetics*. 9(6), pp. 715–723.

Grimwood, J., Gordon, L.A., Olsen, A., *et al.* (2004) The DNA sequence and biology of human chromosome 19. *Nature*. 428(6982), pp. 529–535.

doi:10.1038/nature02399.
Gu, W., Zhang, F. and Lupski, J. (2008) Mechanisms for human genomic rearrangements. *PathoGenetics*. 1, pp. 4. doi:10.1186/1755-8417-1-4.
Habibi, M., Mirfakhraie, R., Khani, M., *et al.* (2017) Genetic variations in UGT2B28, UGT2B17, UGT2B15 genes and the risk of prostate cancer: A case-control study. *Gene*. 634, pp. 47–52. doi:https://doi.org/10.1016/j.gene.2017.08.038.
Handsaker, R.E., Van Doren, V., Berman, J.R., *et al.* (2015) Large multiallelic copy number variations in humans. *Nature Genetics*. 47(3), pp. 296–303. doi:10.1038/ng.3200.
Hastings, P.J., Ira, G. and Lupski, J.R. (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics*. 2009/01/30. 5(1), pp. 1–9. doi:10.1371/journal.pgen.1000327.
He, Y., Hoskins, J.M. and McLeod, H.L. (2011) Copy number variants in pharmacogenetic genes. *Trends in Molecular Medicine*. 17(5), pp. 244–251. doi:10.1016/j.molmed.2011.01.007.
Hebbring, S.J., Adjei, A.A., Baer, J.L., *et al.* (2007) Human SULT1A1 gene: copy number differences and functional implications. *Human Molecular Genetics*. 16(5), pp. 463–470. doi:10.1093/hmg/ddl468.
HIV.gov (2021) *Glossary of HIV/AIDS-Related Terms*. Available at: https://clinicalinfo.hiv.gov/sites/default/files/glossary/Glossary-English_HIVinfo.pdf (Accessed: 30 April 2021).
Hovelson, D.H., Xue, Z., Zawistowski, M., *et al.* (2017) Characterization of ADME gene variation in 21 populations by exome sequencing. *Pharmacogenetics and Genomics*. 27(3), pp. 89–100. doi:10.1097/FPC.0000000000000260.
Huddleston, J., Chaisson, M.J.P., Steinberg, K.M., *et al.* (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*. 27(5), pp. 677–685. doi:10.1101/gr.214007.116.
Ilett, K.F., Ethell, B.T., Maggs, J.L., *et al.* (2002) Glucuronidation of dihydroartemisinin in vivo and by human liver microsomes and expressed UDP-glucuronosyltransferases. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*. 30(9), pp. 1005–1012. doi:10.1124/dmd.30.9.1005.
Ingelman-Sundberg, M. (2004) Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. *Trends in Pharmacological Sciences*. 25(4), pp. 193–200. doi:https://doi.org/10.1016/j.tips.2004.02.007.
Ingelman-Sundberg, M. (2005) Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *The Pharmacogenomics Journal*. 5(1), pp. 6–13. doi:10.1038/sj.tpj.6500285.
Ionita-laza, I., Rogers, A.J., Lange, C., *et al.* (2010) Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*. 93(1), pp. 22–26. doi:10.1016/j.ygeno.2008.08.012.
Ishiguro, A., Kubota, T., Ishikawa, H., *et al.* (2004) Metabolic activity of dextromethorphan O-demethylation in healthy Japanese volunteers carrying duplicated CYP2D6 genes: duplicated allele of CYP2D6*10 does not increase CYP2D6 metabolic activity. *International Journal of Clinical Chemistry*. 344(1–2), pp. 201–204. doi:10.1016/j.cccn.2004.03.002.
Ito, T., Tsuji, M., Mori, Y., *et al.* (2015) Effect of CYP2A6*4 genetic polymorphisms on smoking behaviours and nicotine dependence in a general population of Japanese men. *Fukushima Journal of Medical Science*. 61(2), pp. 125–130. doi:10.5387/fms.2015-14.
Jarvis, J.P., Peter, A.P. and Shaman, J.A. (2019) Consequences of CYP2D6 Copy-Number Variation for Pharmacogenomics in Psychiatry. *Frontiers in Psychiatry*.

10, pp. 432. doi:10.3389/fpsyt.2019.00432.
Jiang, Y., Wang, Y. and Brudno, M. (2012) PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*. 28(20), pp. 2576–2583. doi:10.1093/bioinformatics/bts484.
Johansson, I. and Ingelman-Sundberg, M. (2009) CNVs of human genes and their implication in pharmacogenetics. *Cytogenetic and Genome Research*. 123(1–4), pp. 195–204. doi:10.1159/000184709.
Kalow, W., Tang, B.K. and Endrenyi, L. (1998) Hypothesis: comparisons of inter- and intra-individual variations can substitute for twin studies in drug research. *Pharmacogenetics*. 8(4), pp. 283–289. doi:10.1097/00008571-199808000-00001.
Kamataki, T., Fujieda, M., Kiyotani, K., *et al.* (2005) Genetic polymorphism of CYP2A6 as one of the potential determinants of tobacco-related cancer risk. *Biochemical and Biophysical Research communications*. 338(1), pp. 306–310. doi:10.1016/j.bbrc.2005.08.268.
Kassahun, K., McIntosh, I., Cui, D., *et al.* (2007) Metabolism and disposition in humans of raltegravir (MK-0518), an anti-AIDS drug targeting the human immunodeficiency virus 1 integrase enzyme. *Drug metabolism and Disposition: the Biological Fate of Chemicals*. 35(9), pp. 1657–1663. doi:10.1124/dmd.107.016196.
Kenmotsu, H. and Tanigawara, Y. (2015) Pharmacokinetics, dynamics and toxicity of docetaxel: Why the Japanese dose differs from the Western dose. *Cancer Science*. 106(5), pp. 497–504. doi:10.1111/cas.12647.
Kis, O., Sankaran-Walters, S., Hoque, M.T., *et al.* (2016) HIV-1 Alters Intestinal Expression of Drug Transporters and Metabolic Enzymes: Implications for Antiretroviral Drug Disposition. *Antimicrobial Agents and Chemotherapy*. 60(5), pp. 2771–2781. doi:10.1128/AAC.02278-15.
Klambauer, G., Schwarzbauer, K., Mayr, A., *et al.* (2012) cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research*. 40(9), pp. 1–14. doi:10.1093/nar/gks003.
Koboldt, D.C., Larson, D.E., Chen, K., *et al.* (2012) Massively parallel sequencing approaches for characterization of structural variation. *Methods in Molecular Biology.* 838, pp. 369–384. doi:10.1007/978-1-61779-507-7_18.
Korbel, J.O., Abyzov, A., Mu, X.J., *et al.* (2009) PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*. 10(2), pp. 1–14. doi:10.1186/gb-2009-10-2-r23.
Korhonova, M., Doricakova, A. and Dvorak, Z. (2015) Optical Isomers of Atorvastatin, Rosuvastatin and Fluvastatin Enantiospecifically Activate Pregnane X Receptor PXR and Induce CYP2A6, CYP2B6 and CYP3A4 in Human Hepatocytes. *PLoS One*. 10(9), pp. 1–18. doi:10.1371/journal.pone.0137720.
Laäs, D.J. and Naidoo, M. (2018) An evaluation of warfarin use at an urban district-level hospital in KwaZulu-Natal Province, South Africa. *South African Medical Journal.* 108(12), pp. 1046–1050. doi: 10.7196/SAMJ.2018.v108i12.13256.
Layer, R.M., Chiang, C., Quinlan, A.R., *et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*. 15(6), pp. 84. doi:10.1186/gb-2014-15-6-r84.
Lee, J.H. and Jeon, J.T. (2008) Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels. *Cytogenetic and Genome Research*. 123(1–4), pp. 333–342. doi:10.1159/000184725.
Lentz, F., Tran, A., Rey, E., *et al.* (2005) Pharmacogenomics of fluorouracil, irinotecan, and oxaliplatin in hepatic metastases of colorectal cancer: Clinical implications.

American Journal of PharmacoGenomics. 5(1), pp. 21–33. doi:10.2165/00129785-200505010-00002.
Li, Q. and Weina, P. (2010) Artesunate: The Best Drug in the Treatment of Severe and Complicated Malaria. *Pharmaceuticals*. 3(7), pp. 2322–2332. doi:10.3390/ph3072322.
Lin, C.F., Naj, A.C. and Wang, L.S. (2013) Analyzing copy number variation using SNP array data: protocols for calling CNV and association tests. *Current Protocols in Human Genetics*. 79, pp. 1–17. doi:10.1002/0471142905.hg0127s79.
MacDonald, J.R., Ziman, R., Yuen, R.K.C., *et al.* (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*. 42, pp. 986–992. doi:10.1093/nar/gkt958.
Maheshwari, R., Kuche, K., Mane, A., *et al.* (2018) 'Manipulation of Physiological Processes for Pharmaceutical Product Development' in Tekade, R.K. (ed.) *Dosage Form Design Consideration*. Massachusetts: Academic Press, pp. 701–729.
Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., *et al.* (2019) Structural variant calling: the long and the short of it. *Genome Biology*. 20(1), pp. 246. doi:10.1186/s13059-019-1828-7.
Manosuthi, W., Wiboonchutikul, S. and Sungkanuparph, S. (2016) Integrated therapy for HIV and tuberculosis. *AIDS Research and Therapy*. 13(1), pp. 22. doi:10.1186/s12981-016-0106-y.
Mantere, T., Kersten, S. and Hoischen, A. (2019) Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in Genetics*. 10, pp. 426. doi:10.3389/fgene.2019.00426.
Martis, S., Mei, H., Vijzelaar, R., *et al.* (2013) Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. *The Pharmacogenomics Journal*. 13(6), pp. 558–566. doi:10.1038/tpj.2012.48.
Matimba, A., Dhoro, M. and Dandara, C. (2016) Is there a role of pharmacogenomics in Africa. *Global Health, Epidemiology and Genomics*. 1, pp. 9. doi:10.1017/gheg.2016.4.
Mayosi, B.M., Flisher, A.J., Lalloo, U.G., *et al.* (2009) The burden of non-communicable diseases in South Africa. *Lancet*. 374(9693), pp. 934–947. doi:10.1016/S0140-6736(09)61087-4.
McCarroll, S.A., Hadnott, T.N., Perry, G.H., *et al.* (2006) Common deletion polymorphisms in the human genome. *Nature Genetics*. 38(1), pp. 86–92. doi:10.1038/ng1696.
Mehrotra, N., Gupta, M., Kovar, A., *et al.* (2007) The role of pharmacokinetics and pharmacodynamics in phosphodiesterase-5 inhibitor therapy. *International Journal of Impotence Research*. 19(3), pp. 253–264. doi:10.1038/sj.ijir.3901522.
Meintjes, G., Moorhouse, M.A., Carmona, S., *et al.* (2017) Adult antiretroviral therapy guidelines 2017. *Southern African journal of HIV medicine*. 18(1), pp. 776. doi:10.4102/sajhivmed.v18i1.776.
Ménard, V., Eap, O., Harvey, M., *et al.* (2009) Copy-number variations (CNVs) of the human sex steroid metabolizing genes UGT2B17 and UGT2B28 and their associations with a UGT2B15 functional polymorphism. *Human Mutation*. 30(9), pp. 1310–1319. doi:https://doi.org/10.1002/humu.21054.
Meyerson, M., Gabriel, S. and Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*. 11(10), pp. 685–696. doi:10.1038/nrg2841.
Mills, R.E., Walter, K., Stewart, C., *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*. 470, pp. 59–65.

doi:10.1038/nature09708.Mapping.
Mitchell, C., Gregersen, N. and Krause, A. (2011) Novel CYP2C9 and VKORC1 gene variants associated with warfarin dosage variability in the South African black population. *Pharmacogenomics*. 12(7), pp. 953–63. doi:10.2217/pgs.11.36.
Mroziewicz, M. and Tyndale, R. (2010) Pharmacogenetics: a tool for identifying genetic factors in drug dependence and response to treatment. *Addicition Science and Clinical Practice*. 5(2), pp. 17–29.
Mukonzo, J.K., Owen, J.S., Ogwal-Okeng, J., *et al.* (2014) Pharmacogenetic-based efavirenz dose modification: suggestions for an African population and the different CYP2B6 genotypes. *PLoS One*. 9(1), pp. 1–8. doi:10.1371/journal.pone.0086919.
Navaratnam, V., Mansor, S.M., Sit, N.W., *et al.* (2000) Pharmacokinetics of artemisinin-type compounds. *Clinical Pharmacokinetics*. 39(4), pp. 255–270. doi:10.2165/00003088-200039040-00002.
Ndagije, H.B., Nambasa, V., Manirakiza, L., *et al.* (2018) The Burden of Adverse Drug Reactions Due to Artemisinin-Based Antimalarial Treatment in Selected Ugandan Health Facilities: An Active Follow-Up Study. *Drug Safety*. 41(8), pp. 753–765. doi:10.1007/s40264-018-0659-x.
Nguyen, D.Q., Webber, C. and Ponting, C.P. (2006) Bias of Selection on Human Copy-Number Variants. *PLoS Genetics*. 2(2), pp. 197–207. doi: 10.1371/journal.pgen.0020020
Nijkamp, J.F., Van Den Broek, M.A., Geertman, J.M.A., *et al.* (2012) De novo detection of copy number variation by co-assembly. *Bioinformatics*. 28(24), pp. 3195–3202. doi:10.1093/bioinformatics/bts601.
Nyangiri, O.A., Noyes, H., Mulindwa, J., *et al.* (2020) Copy number variation in human genomes from three major ethno-linguistic groups in Africa. *BMC Genomics*. 21(1), pp. 289. doi:10.1186/s12864-020-6669-y.
Nyirenda, M.J. (2016) Non-communicable diseases in sub-Saharan Africa: Understanding the drivers of the epidemic to inform intervention strategies. *International Health*. 8(3), pp. 157–158. doi:10.1093/inthealth/ihw021.
Ogutu, B. (2013) Artemether and lumefantrine for the treatment of uncomplicated Plasmodium falciparum malaria in sub-Saharan Africa. *Expert Opinion on Pharmacotherapy*. 14(5), pp. 643–654. doi:10.1517/14656566.2013.771167.
Olsson, A.G., McTaggart, F. and Raza, A. (2002) Rosuvastatin: a highly effective new HMG-CoA reductase inhibitor. *Cardiovascular Drug Reviews*. 20(4), pp. 303–328. doi:10.1111/j.1527-3466.2002.tb00099.x.
Pacey, S., Workman, P. and Sarker, D. (2011) 'Pharmacokinetics and Pharmacodynamics in Drug Development' in Schwab, M. (ed.) *Encyclopedia of Cancer.* Berlin: Springer, pp. 2845–2848. doi:10.1007/978-3-642-16483-5_4501.
Pacific Biosciences (2017) *pbsv*. Available at: https://github.com/PacificBiosciences/pbsv (Accessed: 3 May 2021).
Paul, A. (2019) 'Drug Absorption and Bioavailability' in Raj, G.M. and Raveendran, R. (eds.) *Introduction to Basics of Pharmacology and Toxicology*. Singapore: Springer, pp. 81–88. doi:10.1007/978-981-32-9779-1_5.
PharmaADME.org (2021) *PharmaADME.org*. Available at: http://pharmaadme.org. (Accessed: 15 June 2019)
Phillips, A.N., Venter, F., Havlir, D., *et al.* (2019) Risks and benefits of dolutegravir-based antiretroviral drug regimens in sub-Saharan Africa: a modelling study. *The Lancet.* 6(2), pp. 116–127. doi:10.1016/S2352-3018(18)30317-5.
Piacentini, S., Polimanti, R., Porreca, F., *et al.* (2011) GSTT1 and GSTM1 gene polymorphisms in European and African populations. *Molecular Biology Reports*. 38(2), pp. 1225–1230. doi:10.1007/s11033-010-0221-0.

Pinto, D., Darvishi, K., Shi, X., *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*. 29(6), pp. 512–520. doi:10.1038/nbt.1852.

Pirooznia, M., Goes, F. and Zandi, P.P. (2015) Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics*. 6(138), pp. 1–9. doi:10.3389/fgene.2015.00138.

Pös, O., Radvanszky, J., Buglyó, G., *et al*. (2021) DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal*. 44(5), pp. 548–559. doi: 10.1016/j.bj.2021.02.003.

Rajman, I., Knapp, L. and Hanna, I. (2020) Genetic Diversity in Drug Transporters: Impact in African Populations. *Clinical and Translational Science*. 13(5), pp. 848–860. doi:10.1111/cts.12769.

Ramírez, B., Niño-Orrego, M.J., Cárdenas, D., *et al.* (2019) Copy number variation profiling in pharmacogenetics CYP-450 and GST genes in Colombian population. *BMC Medical Genomics*. 12(1), pp. 110. doi:10.1186/s12920-019-0556-x.

Rausch, T., Zichner, T., Schlattl, A., *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 28(18), pp. 333–339. doi:10.1093/bioinformatics/bts378.

Rautio, J., Meanwell, N.A., Di, L., *et al.* (2018) The expanding role of prodrugs in contemporary drug design and development. *Nature Reviews Drug Discovery*. 17(8), pp. 559–587. doi:10.1038/nrd.2018.46.

Redon, R., Ishikawa, S., Fitch, K.R., *et al.* (2006) Global variation in copy number in the human genome. *Nature*. 444(7118), pp. 444–454. doi:10.1038/nature05329.

Rotimi, C., Abayomi, A., Abimiku, A., *et al.* (2014) Research capacity. Enabling the genomic revolution in Africa. *Science*. 344(6190), pp. 1346–1348. doi:10.1126/science.1251546.

Saini, R., Saini, S. and Sugandha, R. (2010) Pharmacogenetics: The future medicine. *Journal of Advanced Pharmaceutical Technology & Research*. 1(4), pp. 423–424. doi:10.4103/0110-5558.76443.

Santos, M., Niemi, M., Hiratsuka, M., *et al.* (2018) Novel copy-number variations in pharmacogenes contribute to interindividual differences in drug pharmacokinetics. *Genetics in Medicine*. 20(6), pp. 622–629. doi:10.1038/gim.2017.156.

Sarkar, S. and Ganguly, A. (2016) Current Overview of Anti-Tuberculosis Drugs: Metabolism and Toxicities. *Mycobacterial Diseases*. 6(2), pp. 1–6. doi:10.4172/2161-1068.1000209.

Scherrmann, J.M. (2009) Transporters in absorption, distribution, and elimination. *Chemistry & Biodiversity*. 6(11), pp. 1933–1942. doi:10.1002/cbdv.200900171.

Scripture, C.D. and Pieper, J.A. (2001) Clinical pharmacokinetics of fluvastatin. *Clinical Pharmacokinetics*. 40(4), pp. 263–281. doi:10.2165/00003088-200140040-00003.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., *et al.* (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*. 15(6), pp. 461–468. doi:10.1038/s41592-018-0001-7.

Semakula, J., Kisa, G., Mouton, H., *et al.* (2021) Anticoagulation in sub-Saharan Africa: Are Direct Oral Anticoagulants the answer? A review of lessons learnt from warfarin. *British Journal of Clinical Pharmacology*. 87(10), pp. 3699–3705. doi:10.1111/bcp.14796.

Semakula, J.R., Mouton, J.P., Jorgensen, A., *et al.* (2020) A cross-sectional evaluation of five warfarin anticoagulation services in Uganda and South Africa. *PLoS One.* 15(1), pp. 1–9. doi:10.1371/journal.pone.0227458.

Shaikh, T.H., Gai, X., Perin, J.C., *et al.* (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and

research applications. *Genome Research*. 19(9), pp. 1682–1690. doi:10.1101/gr.083501.108.
Sim, S.C., Kacevska, M. and Ingelman-Sundberg, M. (2013) Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. *The Pharmacogenomics Journal*. 13(1), pp. 1–11. doi:10.1038/tpj.2012.45.
Sistonen, J., Sajantila, A., Lao, O., *et al.* (2007) CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. *Pharmacogenetics and Genomics*. 17(2), pp. 93–101. doi:10.1097/01.fpc.0000239974.69464.f2.
Sloan, D.J., McCallum, A.D., Schipani, A., *et al.* (2017) Genetic Determinants of the Pharmacokinetic Variability of Rifampin in Malawian Adults with Pulmonary Tuberculosis. *Antimicrobial Agents and Chemotherapy*. 61(7), pp. 1–9. doi:10.1128/AAC.00210-17.
de Smith, A.J., Tsalenko, A., Sampas, N., *et al.* (2007) Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Human Molecular Genetics*. 16(23), pp. 2783–2794. doi:10.1093/hmg/ddm208.
de Smith, A.J., Walters, R.G., Froguel, P., *et al.* (2009) Human genes involved in copy number variation: Mechanisms of origin, functional effects and implications for disease. *Cytogenetic and Genome Research*. 123(4), pp. 17–26. doi:10.1159/000184688.
Soko, N.D., Masimirembwa, C. and Dandara, C. (2018) Rosuvastatin pharmacogenetics in African populations. *Pharmacogenomics*. 19(18), pp. 1373–1375. doi:10.2217/pgs-2018-0168.
Spratlin, J. and Sawyer, M.B. (2007) Pharmacogenetics of paclitaxel metabolism. *Critical Reviews in Oncology/Hematology*. 61(3), pp. 222–229. doi:https://doi.org/10.1016/j.critrevonc.2006.09.006.
Stancu, M.C., van Roosmalen, M.J., Renkens, I., *et al.* (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications*. 8(1), pp. 1326. doi:10.1038/s41467-017-01343-4.
Steen, V.M., Andreassen, O.A., Daly, A.K., *et al.* (1995) Detection of the poor metabolizer-associated CYP2D6(D) gene deletion allele by long-PCR technology. *Pharmacogenetics*. 5(4), pp. 215–223. doi:10.1097/00008571-199508000-00005.
Stingl, J.C., Brockmöller, J. and Viviani, R. (2013) Genetic variability of drug-metabolizing enzymes: the dual impact on psychiatric therapy and regulation of brain function. *Molecular Psychiatry*. 18(3), pp. 273–287. doi:10.1038/mp.2012.42.
Stranger, B.E., Forrest, M.S., Dunning, M., *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 315(5813), pp. 848–853. doi:10.1126/science.1136678.
Takahashi, H. and Echizen, H. (2003) Pharmacogenetics of CYP2C9 and interindividual variability in anticoagulant response to warfarin. *The Pharmacogenomics Journal*. 3(4), pp. 202–214. doi:10.1038/sj.tpj.6500182.
Teo, S.M., Pawitan, Y., Ku, C.S., *et al.* (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*. 28(21), pp. 2711–2718. doi:10.1093/bioinformatics/bts535.
The Cancer Atlas (2018) *Sub-Saharan Africa*. Available at: https://canceratlas.cancer.org/the-burden/sub-saharan-africa/ (Accessed: 2 May 2021).
del Tredici, A.L., Malhotra, A., Dedek, M., *et al.* (2018) Frequency of CYP2D6 Alleles

Including Structural Variants in the United States. *Frontiers in pharmacology*. 9, pp. 305. doi:10.3389/fphar.2018.00305.
Tremmel, R., Klein, K., Battke, F., *et al.* (2020) Copy number variation profiling in pharmacogenes using panel-based exome resequencing and correlation to human liver expression. *Human Genetics*. 139(2), pp. 137–149. doi:10.1007/s00439-019-02093-7.
Tyndale, R.F. and Sellers, E.M. (2001) Variable CYP2A6-mediated nicotine metabolism alters smoking behavior and risk. *Drug Metabolism and Disposition.* 29(4), pp. 548–552.
Vijzelaar, R., Botton, M.R., Stolk, L., *et al.* (2018) Multi-ethnic SULT1A1 copy number profiling with multiplex ligation-dependent probe amplification. *Pharmacogenomics*. 19(9), pp. 761–770. doi:10.2217/pgs-2018-0047.
Wang, H., Veldink, J.H., Blauw, H., *et al.* (2009) Markov Models for inferring copy number variations from genotype data on Illumina platforms. *Human Heredity*. 68(1), pp. 1–22. doi:10.1159/000210445.
Wang, K., Li, M., Hadley, D., *et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*. 17(11), pp. 1665–1674. doi:10.1101/gr.6861907.
Wang, P., Pradhan, K., Zhong, X.-B., *et al.* (2016) Isoniazid metabolism and hepatotoxicity. *Acta Pharmaceutica Sinica B*. 6(5), pp. 384–392. doi:10.1016/j.apsb.2016.07.014.
Warnich, L., Drögemöller, B.I., Pepper, M.S., *et al.* (2011) Pharmacogenomic Research in South Africa: Lessons Learned and Future Opportunities in the Rainbow Nation. *Current Pharmacogenomics and Personalized Medicine*. 9(3), pp. 191–207. doi:10.2174/187569211796957575.
Wasunna, B., Zurovac, D., Goodman, C.A., *et al.* (2008) Why don't health workers prescribe ACT? A qualitative study of factors affecting the prescription of artemether-lumefantrine. *Malaria journal*. 7, pp. 29. doi:10.1186/1475-2875-7-29.
Weinshilboum, R. (2003) Inheritance and Drug Response. *New England Journal of Medicine*. 348(6), pp. 529–537. doi:10.1056/NEJMra020021.
Wester, K., Jönsson, A., Spigset, O., *et al.* (2008) Incidence of fatal adverse drug reactions: A population based study. *British Journal of Clinical Pharmacology*. 65, pp. 573–579. doi:10.1111/j.1365-2125.2007.03064.x.
Wilkinson, G.R. (2005) Drug metabolism and variability among patients in drug response. *The New England Journal of Medicine*. 352(21), pp. 2211–2221. doi:10.1056/NEJMra032424.
Williams, J.A., Hyland, R., Jones, B.C., *et al.* (2004) Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metabolism and Disposition.* 32(11), pp. 1201–1208. doi:10.1124/dmd.104.000794.
Williamson, B., Dooley, K.E., Zhang, Y., *et al.* (2013) Induction of Influx and Efflux Transporters and Cytochrome P450 3A4 in Primary Human Hepatocytes by Rifampin, Rifabutin, and Rifapentine. *Antimicrobial Agents and Chemotherapy*. 57(12), pp. 6366–6369. doi:10.1128/AAC.01124-13.
Wilson, W., de Villena, F.P.M, Lyn-Cook, B.D., *et al.* (2004) Characterization of a common deletion polymorphism of the UGT2B17 gene linked to UGT2B15. *Genomics*. 84(4), pp. 707–714. doi:https://doi.org/10.1016/j.ygeno.2004.06.011.
World Health Organization (2018) *Updated recommendations on first-line and second-line antiretroviral regimens and post-exposure prophylaxis and recommendations on early infant diagnosis of HIV*. Available at: https://www.who.int/publications/i/item/WHO-CDS-HIV-18.51 (Accessed: 20 April

2020).
World Health Organization (2019) *The 'World malaria report 2019' at a glance*. Available at: https://www.who.int/news-room/feature-stories/detail/world-malaria-report-2019 (Accessed: 30 April 2020).
World Health Organization (2020a) *Malaria*. Available at: https://www.who.int/news-room/fact-sheets/detail/malaria (Accessed: 30 April 2020).
World Health Organization (2020b) *Tuberculosis*. Available at: https://www.who.int/health-topics/tuberculosis#tab=tab_1 (Accessed: 30 April 2020).
Yan, M.M., Wu, S.S., Ying, Y.Q., *et al.* (2018) Safety assessment of concurrent statin treatment and evaluation of drug interactions in China. *SAGE Open Medicine*. 6, pp. 1–9. doi:10.1177/2050312118798278.
Ye, K., Schulz, M.H., Long, Q., *et al.* (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 25(21), pp. 2865–2871. doi:10.1093/bioinformatics/btp394.
Yoon, S., Xuan, Z., Makarov, V., *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*. 19(9), pp. 1586–1592. doi:10.1101/gr.092981.109.
Yu, X., Kubota, T., Dhakal, I., *et al.* (2013) Copy number variation in sulfotransferase isoform 1A1 (SULT1A1) is significantly associated with enzymatic activity in Japanese subjects. *Pharmacogenomics and Personalized Medicine*. 6, pp. 19–24. doi:10.2147/PGPM.S36579.
Zang, M., Zhu, F., Zhao, L., *et al.* (2014) The effect of UGTs polymorphism on the auto-induction phase II metabolism-mediated pharmacokinetics of dihydroartemisinin in healthy Chinese subjects after oral administration of a fixed combination of dihydroartemisinin-piperaquine. *Malaria Journal*. 13(1), pp. 478. doi:10.1186/1475-2875-13-478.
Zanger, U.M. and Klein, K. (2013) Pharmacogenetics of cytochrome P450 2B6 (CYP2B6): Advances on polymorphisms, mechanisms, and clinical relevance. *Frontiers in Genetics*. 4(24), pp. 1–12. doi:10.3389/fgene.2013.00024.
Zhang, J., Wang, J. and Wu, Y. (2012) An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinformatics*. 13(6), pp. 1–11. doi: 10.1186/1471-2105-13-S6-S6.
Zhang, Z., Cheng, H., Hong, X., *et al.* (2019) EnsembleCNV: an ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data. *Nucleic Acids Research*. 47(7), pp. 1–13. doi:10.1093/nar/gkz068.
Zhang, Z., Lange, K. and Sabatti, C. (2012) Reconstructing DNA copy number by joint segmentation of multiple sequences. *BMC Bioinformatics*. 13, pp. 205. doi:10.1186/1471-2105-13-205.
Zhang, Z.D., Du, J., Lam, H., *et al.* (2011) Identification of genomic indels and structural variations using split reads. *BMC Genomics*. 12, pp. 375. doi:10.1186/1471-2164-12-375.
Zhao, M., Wang, Qingguo, Wang, Quan, *et al.* (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 14(11), pp. 1–16. doi:10.1186/1471-2105-14-S11-S1.
Zhou, B., Ho, S.S., Zhang, X., *et al.* (2017) Whole-genome sequencing analysis of genomic copy number variation (CNV) using low-coverage and paired-end strategies is highly efficient and outperforms array based CNV analysis. *Journal of Medical Genetics*. 55(11), pp. 735–743. doi: 10.1136/jmedgenet-2018-105272.
Zhou, K. and Pearson, E. (2013). Insights from genome-wide association studies of drug response. *Annual Review of Pharmacology and Toxicology*. 53, pp. 299–

310. doi: 10.1146/annurev-pharmtox-011112-140237.
Zhou, S.F. (2009) Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I. *Clinical Pharmacokinetics*. 48(11), pp. 689–723. doi:10.2165/11318030-000000000-00000.
Zhou, X.Y., Hu, X.X., Wang, C.C., *et al.* (2019) Enzymatic Activities of CYP3A4 Allelic Variants on Quinine 3-Hydroxylation In Vitro. *Frontiers in Pharmacology*. 10, pp. 591. doi:10.3389/fphar.2019.00591.
Zhu, H. and Zhou, X. (2020) Statistical methods for SNP heritability estimation and partition: A review. *Computational and Structural Biotechnology Journal*. 18, pp. 1557–1568. doi: 10.1016/j.csbj.2020.06.011.

CHAPTER 2: AN ASSESSMENT OF THE CNV LANDSCAPE IN ADME GENES USING SHORT-READ SEQUENCING APPROACHES.

2. CHAPTER 2: AN ASSESSMENT OF THE CNV LANDSCAPE IN ADME GENES USING SHORT-READ SEQUENCING APPROACHES.

2.1 Introduction

Copy number variation (CNV) is a type of structural variation (SV) that involves a change in the copy number (CN) (deletions of duplications) of a fragment of DNA that is greater than 50 bp in size. This type of genetic variation can have various functional effects ranging from gene amplification or deletion and changes in gene expression to altered splicing and the production of aberrant proteins. While other types of genetic variation such as single nucleotide variants (SNVs) have been extensively studied, CNV has remained widely understudied, to the extent that this type of variation is often omitted in sequence analysis studies (Eggertsson *et al.*, 2019), predominantly due to the difficulties associated with accurately detecting it.

Various techniques for the detection of CNVs have been used over the years, starting with karyotyping and fluorescence *in situ* hybridization (FISH), followed by array-based approaches (aCGH and SNP arrays) and more recently whole exome sequencing (WES) and whole genome sequencing (WGS) (Carter, 2007; Buysse *et al.*, 2009). Each new technology has provided both higher throughput and higher resolution, from karyotyping which only allows for the detection of large-scale changes such as trisomies at the single-sample level, to whole genome sequencing which can be performed on thousands of samples and has much finer resolution, down to the base-pair level (Coutelier *et al.*, 2021).

However, even with these improvements in technology, calling CNVs from short-read sequencing data remains challenging (Gabrielaite *et al.*, 2021). There are four approaches that can be used for calling CNVs from short-read WGS data namely, Read Pair (RP), Read Depth (RD), Split Read (SR) and Assembly (AS) methods, and many of the newer tools have been designed to combine several approaches (Zhao *et al.*, 2013; Pirooznia, Goes and Zandi, 2015). However, each approach has its own limitations and advantages in terms of call size ranges, breakpoint accuracy and false discovery rate, and the results obtained from different callers often differ considerably from one another, leading to a significant lack of reproducibility (Coutelier *et al.*, 2021).

Encouragingly, it has been shown that combining several different tools can improve the results obtained (Coutelier *et al.*, 2021).

An individual's genetic makeup plays a large role in how they respond to drugs, so a given drug may be effective in one group of patients but may be ineffective or result in serious adverse drug reactions (ADRs) in another (Ahmed *et al.*, 2016). Pharmacogenomics is the study of how genetic variation alters an individual's response to drugs, with the aim of using it as a rational means to optimise drug therapy (Saini, Saini and Sugandha, 2010; Crews *et al.*, 2012). In particular, genetic variation within absorption, distribution, metabolism and excretion (ADME) genes explains to a large extent failure to respond or adverse reactions when a standard dose of a drug is administered (Johansson and Ingelman-Sundberg, 2009).

While SNVs within ADME genes have been well-characterised, less work has been carried out to gain an understanding of the CNV landscape within these genes. This is particularly evident in African populations, which remain understudied in general. Therefore, the aim of this work is to assess the CNV landscape in ADME genes in Sub-Saharan African (SSA) populations using several different CNV calling algorithms, namely Genome STRiP (Handsaker *et al.*, 2015), Manta (Chen *et al.*, 2016), Lumpy (Layer *et al.*, 2014), Delly (Rausch *et al.*, 2012) and GATK (Poplin *et al.*, 2018).

2.2 Materials and Methods

2.2.1 Study participants

A total of 953 short-read, high-coverage (30x – 50x), WGSs from individuals from across SSA were analysed in this study. These individuals originated from west, central, east and southern Africa (*Figure 2.1*), and the sequences were obtained from a number of different research studies. The majority of samples originated from west Africa (n = 564), followed by southern Africa (n = 211), east Africa (n = 148) and central Africa (n = 30) (*Figure 2.1*). A summary of the source, country, self-identified ethnolinguistic group, coverage and number of sequences can be seen in *Table 2.1* below. For all sequences, reads were previously aligned to the build 38 reference genome to generate BAM files using BWA-MEM. These BAM files were used as input for the CNV calling tools that will be discussed below.

Table 2.1: Source, country and self-identified ethnolinguistic group of sequences (n = 953) utilized in this study, as well as coverage and number of sequences.

Source of sequences	Country	Ethnolinguistic group	Coverage	Number of sequences
H3Africa Consortium	Benin	Fon	30x	50
	Burkina Faso	Mossi		34
	Botswana	Various		48
	Cameroon	Bamileke, Ngumba & Mundani		26
	Ghana	Kassena		26
	Nigeria	Barom		48
	Zambia	Soli/Chikunda, Tumbuka & Bemba		41
SAHGP	South Africa	Sotho & Xhosa	50x	15
WITS CBRL	South Africa	SE Bantu speaking	30x	40
AWI-Gen	South Africa	SE Bantu speaking	40x	100
SGDP	Various	Various	40x	25
1KGP	Kenya	Luhya	30x	99
	Nigeria	Esan		99
		Yoruba		107
	Sierra Leone	Mende		85
	The Gambia	Mandinka		110
			TOTAL	**953**

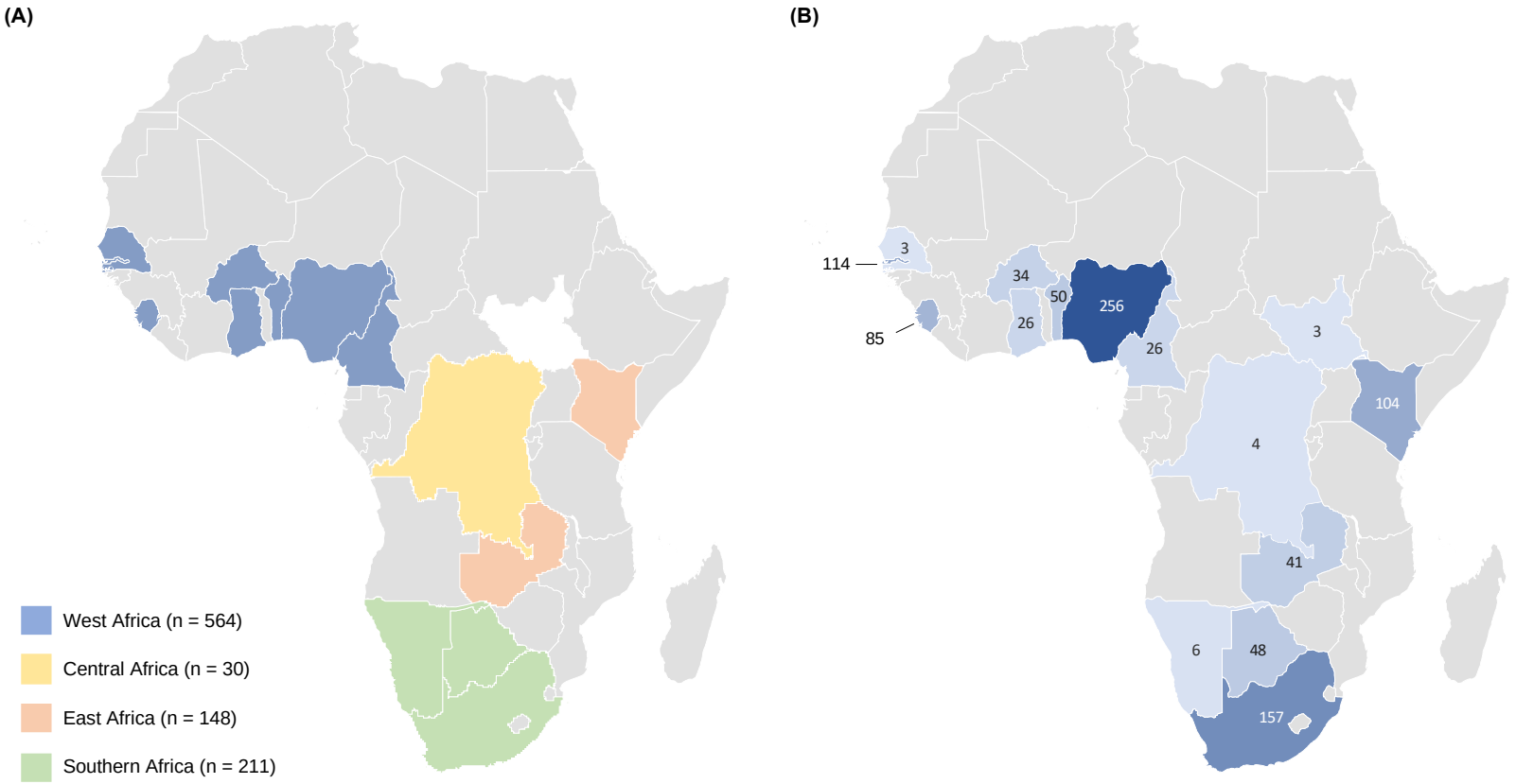


Figure 2.1: (A) Number of samples across west, central, east and southern Africa; (B) Number of samples per country.

2.2.2 Selection of ADME genes

The 299 ADME genes from PharmaADME were included in this study (PharmaADME.org, 2021). This consisted of 32 core ADME genes (*Table A1*) and 267 extended ADME genes (*Table A2*). These genes have been further separated into four categories (*Figure 2.2*): (1) Phase I metabolising enzymes which convert the parent drug into more active metabolites; (2) Phase II metabolising enzymes which convert the parent drug into more inactive metabolites; (3) transporters which are involved in the transport of the drug into and out of the cells and (4) modifiers which are able to alter either the gene expression of other ADME genes or the biochemistry of ADME enzymes (PharmaADME.org, 2021). A table containing the core and extended ADME genes as well as which category they belong to can be seen in Appendix A (*Table A1; Table A2*). The genomic locations of the selected ADME genes were obtained using Ensembl's built in BioMart tool (v104) (Howe *et al.*, 2021).

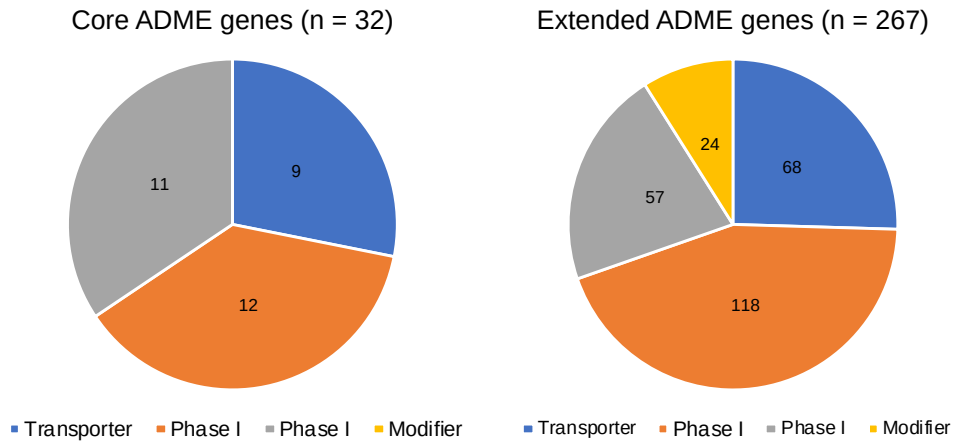


Figure 2.2: Breakdown of the number of Phase I metabolising enzymes, Phase II metabolising enzymes, transporters and modifiers that make up the core and extended ADME gene lists.

2.2.3 CNV calling, processing and quality control

Many tools have been developed to call CNVs from short-read WGS data. While some of these tools are able to call various types of SV, only deletions and duplications were considered as this is the definition of a CNV. As each tool has its own limitations and advantages, the use of a variety of CNV calling algorithms can generate a more comprehensive dataset. In this study CNV calling was performed using several tools

including Genome STRiP, Manta, Delly, Lumpy and GATK (*Figure 2.3*). These tools were chosen as they make use of different combinations of approaches for calling CNVs (*Figure 2.3*). Manta, Delly and Lumpy were included as they have been shown to have high levels of precision and recall (Kosugi *et al.*, 2019), while Genome STRiP was included as it calls CNVs across multiple samples, which is a different approach compared to all of the other tools.

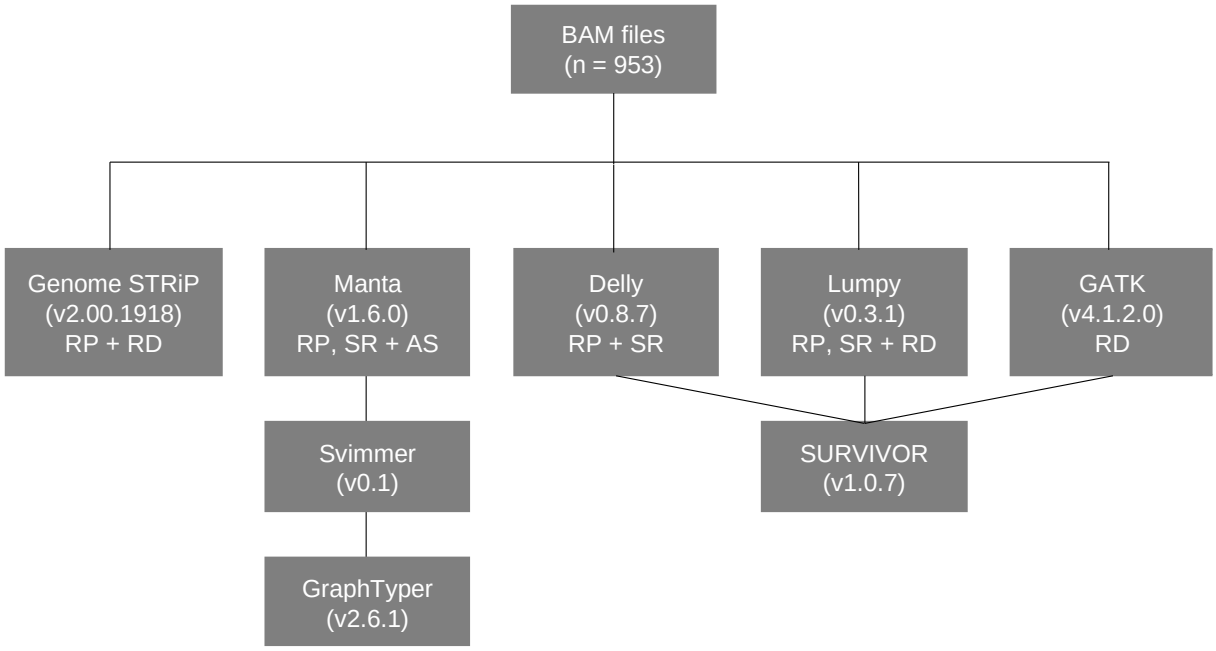


Figure 2.3: CNV calling workflow including the approaches used by each tool to call CNVs. 953 BAM files were run individually through Manta, Delly, Lumpy and GATK and 919 BAM files were run jointly though Genome STRiP.

2.2.3.1 Genome STRiP

Genome STRiP is a suite of tools for discovering and genotyping CNVs from WGS data using a RP and RD approach. Unlike the majority of other tools used for calling CNVs, Genome STRiP is designed to detect shared variation using sequencing data from multiple individuals. A minimum of 20 to 30 genomes are required for acceptable results; however, the method gains power across genomes and processing more genomes improves the accuracy and provides better results. Genome STRiP consists of two main pipelines, namely SVPreprocess and CNVDiscovery. The SVPreprocess pipeline computes summary metadata for the input dataset and is a pre-requisite for all downstream pipelines. The CNVDiscovery pipeline is used for the discovery and genotyping of CNVs. This pipeline is a large and complex pipeline consisting of 12 stages (*Figure 2.4*) and processing within several of the stages is further parallelized

by chromosome. Genome STRiP also has an SVGenotyper pipeline to genotype variants as well as an SVAnnotator pipeline which annotations VCF files containing SV records. A limitation of Genome STRiP is that it is unable to call small CNVs less than 1 000 bp in size.



Figure 2.4: Description of the 12 stages of Genome STRiP's CNVDiscovery pipeline.

Running Genome STRiP

Genome STRiP (svtoolkit_2.00.1918) (Handsaker *et al.*, 2015) uses the Queue workflow manager to implement the analysis pathways. While Queue manages processing steps and dependencies, it does not directly execute the processing steps itself. Instead, the processes are scheduled to run on a compute cluster through a workflow manager. Queue has built-in support for Load Sharing Facility (LSF) or Sun Grid Engine (SGE) workflow managers; however, the compute cluster available for this project uses the Slurm workflow manager. Therefore, a Distributed Resource Management Application API (DRMAA), which acts as an interface between Queue and Slurm, was required.

A total of 34 samples were excluded from this analysis for various reasons. Three samples from South Africa were excluded as the sex of the individual was not available and is required as an input by Genome STRiP. Additionally, 31 samples (19 Cameroon; 11 Botswana; 1 Congo) were defined as substantial outliers and excluded from this analysis as the number of variants per sample was high (> 12 000). These samples may contain a disproportionate number of CNVs as a result of poor sample quality. Therefore, a total of 919 samples were run though the Genome STRiP pipeline.

To minimize batch effects caused by running samples processed with and without PCR together, the dataset was split into three batches, and run through the SVPreprocess and CNVDiscovery pipelines separately (PCR free samples, PCR based samples and AWI-Gen samples). The AWI-Gen data was run in a separate batch as the sequencing approach was unclear; however, it was later confirmed that the AWI-Gen samples were PCR-free. As Genome STRiP could not be successfully run on the X and Y chromosomes, CNV calling was only carried out on the autosomes. SVPreprocess (https://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_qscript_SVPreprocess.html) was run on all of the autosomes at once using default parameters, with a memory limit of 50 GB and a maximum of 75 jobs running in parallel. Due to the large memory requirement of the CNVDiscovery pipeline (https://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_qscript_CNVDiscoveryPipeline.html), it was run on each autosome separately using default parameters, with a memory limit of 100 GB and a maximum of 60 jobs running in parallel. Following CNVDiscovery, the VCF files for each chromosome were merged into a single VCF

using GATK's CatVariants (v4.1.8.1) in a custom bash script which can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%202%20scripts/mergeVCFs.sh).

As sites were discovered in three different batches, we needed to obtain genotypes for the sites in one batch in the other two batches as well. In other words, AWI-Gen sites needed to be genotyped in PCR-free and PCR-based datasets and vice versa. In order to do this we first removed the genotype data from the three VCFs using regular expressions so we were only left with the site information. This resulted in six VCF files, in addition to the three original VCFs. We then ran the SVGenotyper pipeline (https://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_qscript_SVGenotyper.html) which genotypes variants discovered using GenomeSTRiP. Variants were run through the pipeline in batches of 100 and a memory limit of 100 GB was set. The SVGenotyper pipeline produced a stutter where the last few variants of each batch were duplicated in the next batch. A custom python script which can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%202%20scripts/remove_stutter.py) was used to remove these duplicated variants. Lastly, the SVAnnotator pipeline (https://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_main_SVAnnotator.html) was run on the nine VCFs. These nine files were combined into a single merged data frame and reannotated.

Quality control of Genome STRiP data

Quality control of the Genome STRiP data was performed using a pipeline developed by Jakubosky and colleagues (Jakubosky *et al.*, 2020). Variants were filtered based on GSCNQUAL scores, which are per site quality scores generated by Genome STRiP. Different thresholds were used for different variant types, with 2, 12, and 14 set as the minimum GSCNQUAL score required for deletions, multiallelic CNVs (mCNVs) and duplications respectively (Jakubosky *et al.*, 2020). These thresholds were determined by Jakubosky and colleagues to produce a high quality, reproducible dataset (Jakubosky *et al.*, 2020). The deletion threshold is lower than the thresholds for mCNVs and duplications as deletions were found to be more reproducible under less stringent filtering (Jakubosky *et al.*, 2020). Prior to filtering, 61 638 CNVs were identified. This was reduced to 38 973 following filtering based on GSCNQUAL values.

As CNVs were called in three different batches, some variants were redundant across batches (i.e. the same variant was discovered in different datasets) and a variant deduplication procedure was required. The variants discovered in the three batches were intersected using BEDTools, and sites with at least 80% reciprocal overlap were selected. These sites were further categorised into two groups, namely, identical duplicates and non-identical duplicates. Identical duplicates were defined as sites that exactly overlapped (100% overlap), i.e. had the same breakpoints. In these cases, the variant with the highest GSCNQUAL across all batches was selected, and the redundant variant/s were removed from further analysis. Non-identical duplicates were defined as near exact duplicate calls (80–99% reciprocal overlap) that had slightly different breakpoints. Pairs of variants were marked redundant if they (a) had less than 5% of non-mode CN samples discordant, (b) had a Pearson correlation > 0.95 and (c) had a mean difference in CN among non-mode samples and all samples of 0.2. Variants that were identified as redundant were then represented in a graph, with edges between redundant variants weighted based on the average percentage overlap (i.e. the average of the percentage overlap of site 1 with site 2 and the percentage overlap of site 2 with site 1). Iterating through groups of connected components, we chose the pair of sites that had the highest average overlap, and from each pair, the variant with the highest GSCNQUAL was selected and the remaining were pruned. During this deduplication process, 4 406 identical sites were reduced to 1 998 sites and 16 157 non-identical sites were reduced to 6 818 sites.

Genome STRiP tends to report a single CNV as several smaller adjacent or overlapping CNVs (Jakubosky *et al.*, 2020). The abovementioned deduplication process was performed in order to remove near exact duplicate calls made in different batches that differed in breakpoint slightly, but weren't likely to be fragmented calls. To address the fragmentation issue, we stitched adjacent variants whose genotypes correlated because they likely represent a single variant that GenomeSTRiP called as multiple adjacent variants. To do this, we iterated through every adjacent pair of variants on each chromosome and computed the correlation between variants. If the correlation co-efficient was > 0.9 and the variants were less than 30kb apart, the pair of variants were considered valid to be stitched. We then iterated through successive pairs, stitching multiple CNVs together if multiple adjacent pairs passed the criteria for stitching. However, if multiple correlated CNVs in a row are broken up by an uncorrelated pair, the stitching will be broken up. There is therefore a possibility that

rare CNVs may break up the clusters. This process resulting in 11 021 CNVs being combined into 3 539 CNVs.

The newly stitched variants needed to be re-genotyped in the three batches of data separately using the SVGenotyper pipeline, followed by annotation with the SVAnnotator pipeline. We then compared the genotypes of the stitched sites with average of the original genotypes of the constituent sites, and those that did not have a high correlation (average $r < 0.9$ across all constituents) were unstitched. A final VCF file was generated using the following python script which can be seen in the GitHub repository: https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%202%20scripts/StitchSiteFiltering.py. Following the above quality control process, a total of 16 539 CNVs remained.

Quality control issues involving alternate contig regions in the BAM files were identified towards the end of the project. However, as the computational cost of re-generating these BAM files and re-calling CNVs is several hundred thousand CPU hours and takes over four months, all CNVs that were identified within alternate contig regions were excluded from the final dataset. Following this, a total of 15 658 CNVs remained. CNVs falling within ADME gene regions were extracted using BCFtools (v1.10.2), and a total of 78 ADME CNVs were identified. The regeneration of the BAM files and recalling of CNVs to take into account CNVs within alternate contig regions is discussed in Chapter 5.

2.2.3.2 Structural Variation Engine (SVE)

Structural Variation Engine (SVE) is a python-based execution engine for the detection of SVs. It is available as a docker image which simplifies execution, as all of the SV callers and other dependencies do not need to be downloaded onto the compute cluster individually. SVE contains ten SV calling tools namely, Breakdancer, cn.MOPS, CNVnator, Delly, GATK, Genome STRiP, Hydra, Lumpy, Breakseq2 and Tiagra-SV (*Figure 2.5*). Each SV caller produces a VCF file which can be used an input to FusorSV, which merges the SV call sets into a single VCF file using a machine learning approach (*Figure 2.5*).

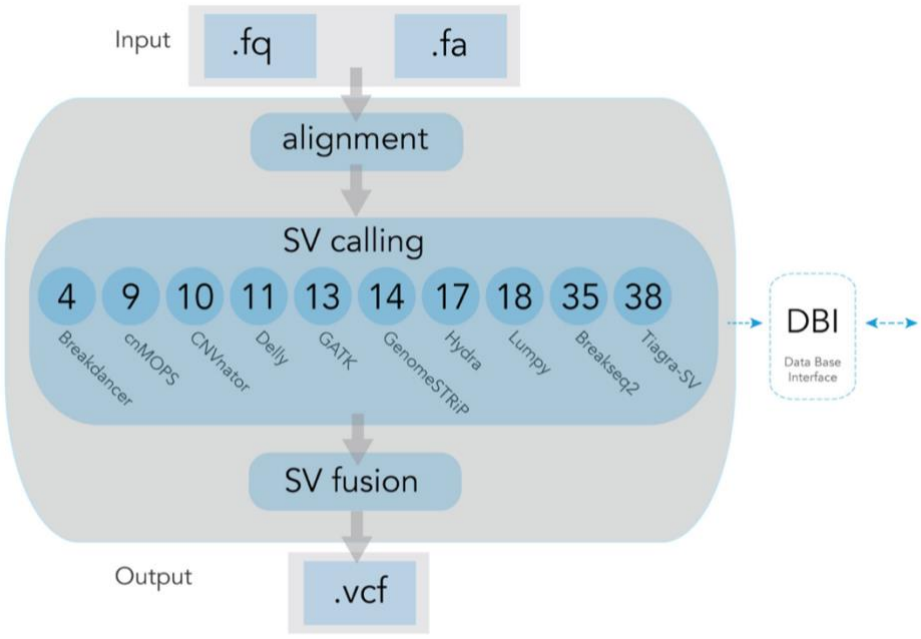


Figure 2.5: SVE workflow (https://github.com/timothyjamesbecker/SVE).

Experimentation of SVE was performed and it was determined that it did not support build 38 aligned data. Subsequently, some of the tools included in SVE were run individually (Delly, Lumpy, GATK) and we performed our own merging.

2.2.3.3 Manta

Manta is an SV discovery tool designed to call SVs from next generation sequencing (NGS) data (Chen *et al.*, 2016). It calls SVs using a RP, SR and AS approach, and has been specifically designed for rapid analysis and large-scale SV calling from WGS data (Chen *et al.*, 2016). Manta can accurately call CNVs between 50 bp–10 Mb in size.

Manta divides the SV discovery process into two steps. Firstly, the genome is scanned to find evidence of possible SVs. This evidence is added onto a graph with edges connecting all of the regions of the genome which have possible breakend associations (Chen *et al.*, 2016). Edges may connect two different regions of the genome to represent long-range association, or an edge may connect to itself to capture a small, local SV association (Chen *et al.*, 2016). Secondly, the graph edges are analysed in order to discover and score SVs. The sub-steps of this process include: (1) inferring SV candidates associated with each edge; (2) assembling breakends of the SVs to base pair resolution; (3) scoring, genotyping and filtration of the SVs and (4) outputting the results to a VCF file (Chen *et al.*, 2016).

As Manta outputs a single VCF file per sample, svimmer (DecodeGenetics, 2019), a tool written for merging SVs discovered using Manta, was used to merge similar SVs from the multiple single sample VCF files. The output from svimmer is a single VCF file that contains all merged SVs. This output can be passed to GraphTyper for genotyping of the merged sites. GraphTyper is a graph-based variant caller that was designed for the genotyping of population-scale short-read datasets using pangenome graphs. It is capable of genotyping various types of genetic variation including SVs (Eggertsson *et al.*, 2019).

Running Manta, svimmer and GraphTyper

Manta, svimmer and GraphTyper were run consecutively using a Nexflow script which, along with the configuration file, can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%202%20scripts/Manta.nf and https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%202%20scripts/Manta.config). Manta (v1.6.0; https://github.com/Illumina/manta) (Chen *et al.*, 2016) was run using default parameters in single sample mode. A maximum of 60 samples were run in parallel, with a memory limit of 5 GB. Merging of variants across all samples was performed using svimmer (v0.1; https://github.com/DecodeGenetics/svimmer) which was run using default parameters and the --threads option set to eight. GraphTyper (v2.6.1; https://github.com/DecodeGenetics/graphtyper) (Eggertsson *et al.*, 2019) was run with default parameters to genotype the merged sites. Each chromosome was run separately due to the large memory requirements (memory limit of 300 GB). The resulting VCF files for each chromosome were merged using VCFtools' vcf-concat function (v0.1.16).

Quality control of Manta data

Variants were filtered for quality using the filter tag 'PASS' as well as the normalised quality score (QD). Variants that did not pass (the default is a QD score of less than 9) were excluded from further analysis. Deletions were further filtered using a stricter cut-off (QD score of less than 12) as was done by Eggertsson and colleagues in order to produce a high-confidence dataset (Eggertsson *et al.*, 2019). Variants were further filtered based on size, with variants that were greater than 10 Mb being excluded, as Manta is unable to call variants larger than this with acceptable accuracy. Furthermore, Manta calls all types of SVs, and as we were only interested in deletions and

duplications, all insertions, inversions, and translocations were filtered out. GraphTyper makes use of two different models for genotyping, namely 'break point' and 'coverage'. It also has an 'aggregated' model that combines both previously mentioned models. Only variants that had been genotyped using the 'aggregated' model were retained for further analysis. A Hardy Weinberg Equilibrium analysis was performed on the remaining variants and excessive heterozygosity was calculated. Variants that had an excessive heterozygosity of $p<1 \times 10^{-15}$ were excluded. Genotypes were set to missing (./.) if their genotype quality (GQ) score was less than 20, and if they had a 'FAIL1' tag for deletions and duplications or a 'FAIL2' or 'FAIL3' tag for duplications. After the low-quality genotypes were set to missing, allele frequency was recalculated using VCFtools (v0.1.16) and variants with an allele frequency of zero (AF = 0) were excluded. Lastly, as discussed in the GenomeSTRiP section, all variants overlapping alternate contig regions were removed. All of the above filtering was performed using BCFtools (v1.10.2) and VCFtools (v0.1.16) as well as some custom python scripts which can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%202%20scripts/AF0.py and https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%202%20scripts/ExHet_filtering.py). After the above quality control, the final Manta dataset consisted of 53 986 CNVs. Variants within ADME gene regions were extracted using BCFtools (v1.10.2), and a total of 288 ADME CNVs remained.

2.2.3.4 Delly, Lumpy & GATK

Delly is an SV calling tool that uses RP and SR approaches for the discovery and genotyping of various types of SV including deletions and duplications (Rausch *et al.*, 2012). Lumpy is a probabilistic framework for the detection of SVs (Layer *et al.*, 2014) that integrates RP, SR and RD approaches (Pirooznia, Goes and Zandi, 2015). A recent study performed by Zhang and colleagues that compared ten different algorithms showed that Lumpy had the best performance in terms of both sensitivity and specificity (Zhang *et al.*, 2019). Lastly, GATK HaplotypeCaller has traditionally been used to call small variants such as SNPs and indels but has recently been extended to CNV calling. This tool makes use of the RD approach for CNV calling. All three of the above tools are able to detect CNVs within the 50 bp–10 Mb size range.

Running Delly, Lumpy and GATK

Delly (v0.8.7; https://github.com/dellytools/delly) (Rausch *et al.*, 2012), Lumpy (v0.3.1; https://github.com/arq5x/lumpy-sv) (Layer *et al.*, 2014) and GATK HaplotypeCaller (v4.1.2.0) (Poplin *et al.*, 2018) were run using default parameters in a single Nextflow script which can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%202%20scripts/Delly_Lumpy_GATK.nf). Due to the large memory requirements, each chromosome was run separately. The individual chromosome VCFs produced by both Lumpy and Delly were combined using BCFtools' concat function (v1.10.2) to produce a single VCF per sample, while the individual chromosome VCFs produced by GATK were merged using GATK's MergeVCF function (v4.1.8.1). All 953 samples were run through these tools, with the exception of ten samples which we excluded from the Lumpy dataset as they failed several attempted analyses using > 500 GB of RAM per chromosome.

Merging results from Delly, Lumpy and GATK

Each of the above mentioned tools produces a single VCF per sample. In order to be able to compare across samples, the CNVs needed to be merged. SURVIVOR's merge function (v1.0.7; https://github.com/fritzsedlazeck/SURVIVOR/) is able to merge and compare SVs within and among samples in order to obtain a consensus call set. This was used to iteratively merge the VCF files produced by Delly, Lumpy and GATK. Firstly, the VCFs produced by each tool were merged for each sample, resulting in a VCF file for each sample containing the merged data from the three tools. Secondly, the individual sample VCFs produced in the first step were merged to generate a single, final VCF containing the SV calls from the three tools across all samples. In both steps, SURVIVOR merge was run with the maximum distance between SVs set to 1 kb, and the minimum SV size set to 30 bp. Furthermore, strand and SV type were considered during the merging.

Quality control of SURVIVOR data

The tools used for variant calling are able to call all SV types; however, as we are only interested in deletions and duplications, all other variant types were filtered out. Deletions and duplications were further filtered based on size. Any variants less than 50 bp or greater than 10 Mb were excluded. CNVs that occurred within ADME regions

were extracted. Lastly, only CNVs with a quality score greater than 1 000 were retained. The above filtering was performed using BCFtools (v1.10.2) as well as SURVIVOR's filter option (v1.0.7; https://github.com/fritzsedlazeck/SURVIVOR/) which allows for CNVs to be filtered based on size. Following the above quality control, the final SURVIVOR dataset consisted of 30 798 CNVs, 158 of which were within ADME genes.

2.2.4 Merging of CNVs from different tools

CNVs were manually inspected and merged based on their start and end positions, CNV type, allele frequency and size. CNVs identified by a single tool were retained as the tools differ in their ability to call different types and size ranges of CNVs. For example, Genome STRiP cannot identify CNVs that are smaller than 1 000 bp, while the other tools can. Additionally, Genome STRiP is the only tool that has the ability to call mCNVs. Where Manta and SURVIVOR calls overlapped, the Manta call was retained. Where Genome STRiP and Manta calls overlapped, the Manta call was retained, except in the case of mCNVs, in which case the Genome STRiP call was retained. Where all three tools overlapped, the Manta call was retained, except in the case of mCNVs, in which case the Genome STRiP call was retained. Manta calls were retained when there was an overlap with other tools as we had the most confidence in this call set.

2.2.5 CNV annotation

Basic Linux commands as well as VCF manipulating tools, such as BCFtools (v1.10.2) and VCFtools (v0.1.16), were used for data analysis. Ensembl's Variant Effect Predictor (VEP) (v104) (Howe *et al.*, 2021) was used to determine which genes were overlapped by CNVs and to assess the CNV consequences. Variants were compared to those in the Database of Genomic Variation (DGV) (v107) (MacDonald *et al.*, 2014) using BEDTools intersect (v2.27.1) with 50% reciprocal overlap in order to determine which CNVs were novel. All figures and graphs were generated using Microsoft Excel (v16.54) or R Studio (ggplot2 package) (v1.2.1335).

2.3 Results

2.3.1 Overview of CNVs identified by each tool

2.3.1.1 CNV counts

To provide a comprehensive overview of the CNV landscape in ADME genes, we analysed 953 high-coverage WGSs using various different CNV calling tools to identify deletions and duplications. An overview of the CNVs identified by each tool can be seen in *Table 2.2* below.

Table 2.2: Count, length and frequency of ADME deletions, duplications and mCNVs identified using Genome STRiP, Manta as well as Delly, Lumpy and GATK (SURVIVOR dataset).

Dataset	CNV category	Count	Average Length (kb)	AF < 1%	1% ≤ AF ≤ 5%	AF > 5%
Genome STRiP	Deletions	52 (66.7%)	10.1	43 (82.7%)	5 (9.6%)	4 (7.7%)
	Duplications	17 (21.8%)	19.6	10 (58.8%)	5 (23.5%)	3 (17.6%)
	mCNVs	9 (11.5%)	4.3	2 (22.2%)	4 (44.4%)	3 (33.3%)
	All CNVs	**78**	**11.5**	**55 (70.5%)**	**13 (16.7%)**	**10 (12.8%)**
Manta	Deletions	243 (84.4%)	297.8	112 (50.2%)	41 (16.9%)	80 (32.9%)
	Duplications	45 (15.6%)	62.8	29 (64.4%)	11 (24.4%)	5 (11.1%)
	All CNVs	**288**	**261.1**	**151 (52.4%)**	**52 (18.1%)**	**85 (29.5%)**
SURVIVOR	Deletions	128 (81%)	228.7	65 (50.8%)	21 (16.4%)	42 (32.8%)
	Duplications	30 (19%)	1012.8	15 (50%)	8 (26.7%)	7 (23.3%)
	All CNVs	**158**	**377.6**	**80 (50.6%)**	**29 (18.4%)**	**49 (31%)**

The Genome STRiP dataset consisted of 15 658 CNVs that were identified across the genome, 78 of which were found within ADME genes (*Table 2.2*). The ADME CNVs called by Genome STRiP comprised 52 deletions (66.7%), 17 duplications (21.8%) and nine mCNVs (11.5%) (*Table 2.2*). mCNVs are defined as regions that are deleted in some individuals and duplicated in others (mixed mCNV), or where individuals all have duplications but with different CNs (duplication mCNV). Of the nine mCNVs detected in ADME genes, eight were mixed mCNVs and only one was a duplication mCNV. When assessing the Manta dataset, a total of 53 986 CNVs were identified across the genome, and 288 of these CNVs occurred within ADME genes (*Table 2.2*). The majority of ADME CNVs called by Manta were deletions (n = 243; 84.4%), while duplications only accounted for 15.6% (n = 45) of ADME CNVs (*Table 2.2*). Lastly, Delly, Lumpy and GATK calls were merged using SURVIVOR and will henceforth be referred to as the SURVIVOR dataset. The SURVIVOR dataset consisted of 30 798 CNVs, 158 of which were ADME CNVs. The ADME CNVs in the SURVIVOR dataset consisted of 128 deletions (81%) and 30 duplications (19%) (*Table 2.2*). Genome STRiP is the only tool that was used that is able to call mCNVs and, as such, no mCNVs were identified in the Manta and SURVIVOR dataset. Manta called more ADME CNVs compared to the Genome STRiP and SURVIVOR datasets, and the proportion of deletions discovered by Manta and SURVIVOR where higher than the proportion of deletions discovered by Genome STRiP.

2.3.1.2 Allele frequencies of ADME CNVs

When assessing the allele frequencies of the ADME CNVs in the three datasets, GenomeSTRiP called a higher proportion of rare CNVs (AF < 1%) (70.5%), compared to Manta (52.4%) and SURVIVOR (50.6%) (*Table 2.2*). In terms of low frequency CNVs (1% ≤ AF ≤ 5%), all three datasets had similar proportions (Genome STRiP: 16.7%; Manta: 18.1% and SURVIVOR: 18.4%) (*Table 2.2*). The proportion of common CNVs (AF > 5%) in the Manta and SURVIVOR datasets was 29.5% and 31% respectively (*Table 2.2*). This is higher than the proportion of common CNVs identified by Genome STRiP which was only 12.8% (*Table 2.2*).

The allele frequency of the CNVs was also assessed for deletions and duplications separately. In the Genome STRiP dataset there was a higher proportion of rare deletions (82.7%) compared to rare duplications (58.8%), while for common CNVs, there is a higher proportion of common duplications (17.6%) compared to common

deletions (7.7%) (*Table 2.2*). The opposite is true for the Manta data where there was a higher proportion of rare duplications (64.4%) compared to rare deletions (50.2%), whilst for common CNVs, there was a higher proportion of common deletions (32.9%) compared to common duplications (11.1%) (*Table 2.2*). Additionally, in the SURVIVOR dataset we saw an equal proportion (~50%) of rare deletions and rare duplications, and the proportion of common deletions was slightly higher (32.8%) than that of common duplications (23.3%) (*Table 2.2*).

2.3.1.3 Sizes of ADME CNVs

The smallest ADME CNV detected by Genome STRiP was 1 130 bp in size, while the largest was 64.7 kb. The average CNV size was 11.5 kb and duplications were larger than deletions, with an average size of 19.6 kb compared to 10.1 kb (*Table 2.2*). mCNVs were even smaller, with and average size of only 4.3 kb (*Table 2.2*). The average CNV size for the Manta dataset was 261.1 kb (*Table 2.2*). The smallest CNV identified by Manta was 51 bp in size, and the largest was 9.7 Mb in size. Unlike in the Genome STRiP dataset, deletions were larger than duplications with an average size of 297.8 kb and 62.8 kb respectively (*Table 2.2*). The minimum and maximum CNV size for the SURVIVOR dataset was 55 bp and 8.8 Mb respectively and the average CNV size was 337.6 kb (*Table 2.2*). Similarly to the Genome STRiP dataset and unlike the Manta STRiP dataset, duplications were larger than deletions with an average length of 1 Mb, compared to 228.7 kb for deletions (*Table 2.2*). On average, the CNVs identified by Manta and SURVIVOR were far larger than those identified by Genome STRiP (*Table 2.2*).

When assessing the size ranges of the CNVs identified in each dataset (*Figure 2.6*), it can be seen that Genome STRiP only called CNVs that were between 1 kb and 100 kb in size, with CNVs between 1 kb and 10 kb being the most common (71.8%). Manta and SURVIVOR were able to call smaller CNVs (<1000 bp), which accounted for 51.4% and 41.1% of the total number of CNVs in each dataset respectively (*Figure 2.6*). Additionally, Manta and SURVIVOR called a number of very large CNVs (100 kb–10 Mb) that GenomeSTRiP was not able to call (*Figure 2.6*).

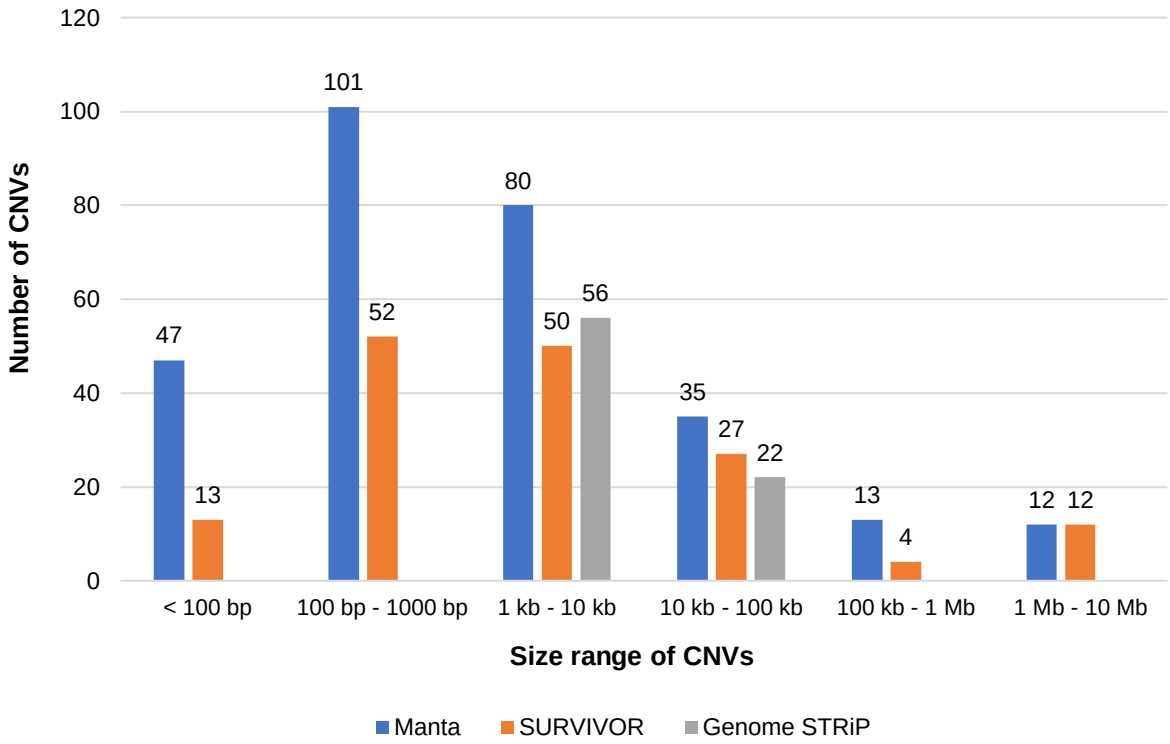


Figure 2.6: Size ranges of ADME CNVs detected by Manta (blue), SURVIVOR (orange) and Genome STRiP (grey).

When the allele frequency was plotted against size for all of the CNVs for each tool (*Figure 2.7A-C*), it can be seen that the majority of small CNVs were common, while larger CNVs tended to be observed at lower allele frequencies. This is with the exception of the SURVIVOR dataset in which several large CNVs were observed at high frequencies (*Figure 2.7C*). When looking only at CNVs that were less than 10 kb in size for each tool (*Figure 2.7D-F*) the same pattern of decreasing allele frequency with increasing size was observed; however, there are several CNVs around the 2 kb and 5 kb size ranges with increased allele frequency, which was observed across all three datasets.

2.3.1.4 Novel ADME CNVs

The CNVs identified in each dataset were compared to the DGV in order to determine if they had been previously reported or were novel. Of the CNVs detected by Genome STRiP, 32.1% (n = 25) did not overlap with any CNVs in the DGV and were classified as novel. Manta identified the highest number of novel CNVs compared to the other tools, with 29.9% (n = 86) of Manta CNVs being classified as novel. Lastly, 26.5% (n = 42) of the SURVIVOR CNVs were classified as novel.

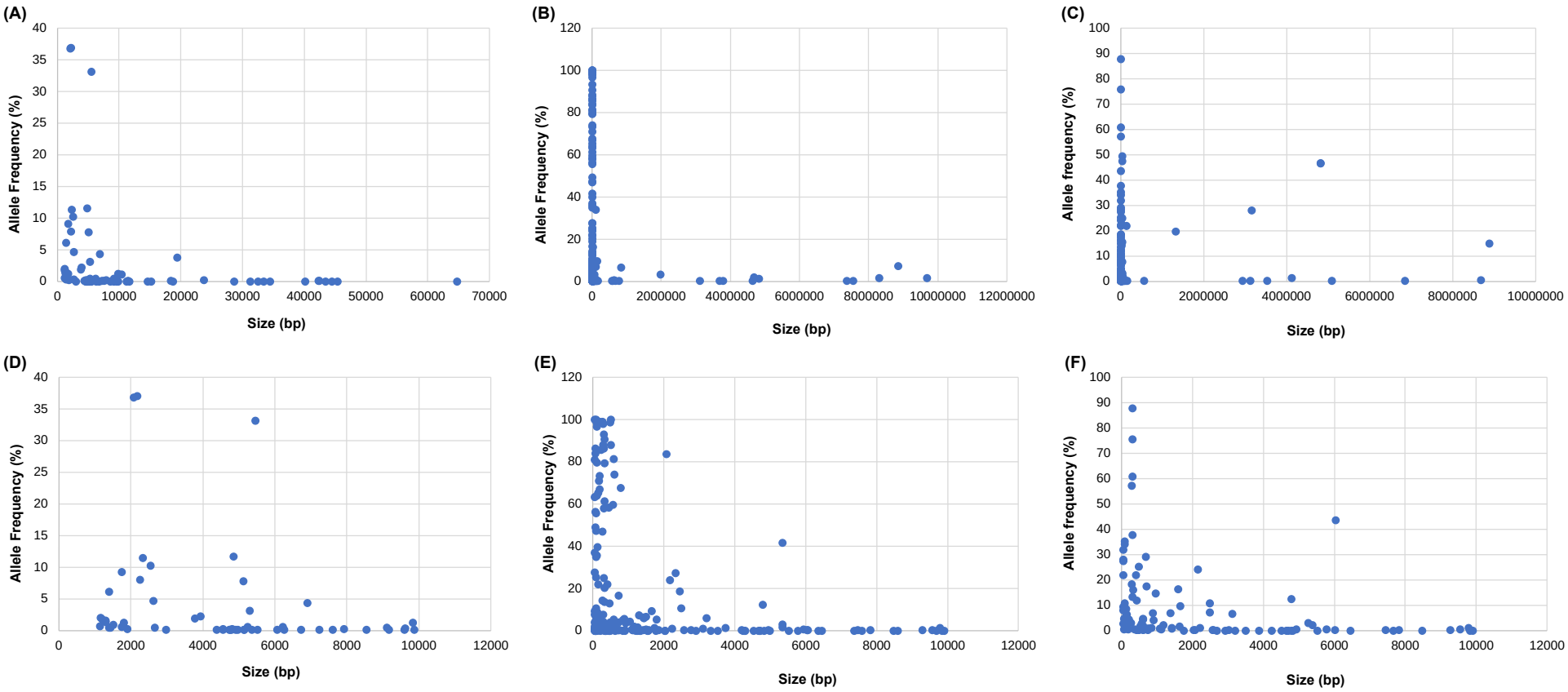


Figure 2.7: Allele frequency versus size for (A) all Genome STRiP CNVs; (B) all Manta CNVs; (C) all SURVIVOR CNVs; (D) GenomeSTRiP CNVs < 10 kb; (E) Manta CNVs < 10 kb and (F) SURVIVOR CNVs < 10 kb.

2.3.2 Overlap between callers

The overlap between the CNVs in the Genome STRiP, Manta and SURVIVOR datasets can be seen in *Figure 2.8A* below. A total of 362 CNVs were identified across all three datasets. Manta had the highest number of unique CNVs (n = 165; 45.6%), followed by SURVIVOR (n = 47; 13%) and Genome STRiP (n = 27; 7.5%). A total of 39 CNVs (10.8%) were common across all three datasets. Manta and Genome STRiP shared an additional 12 CNV (3.3%) and Manta and SURVIVOR shared an additional 72 CNV (19.9%). Interestingly, of the 111 CNVs shared by Manta and SURVIVOR datasets, 62.2% (n = 69) of the CNVs had either identical breakpoints or nearly identical breakpoints within 5 bp of each other.

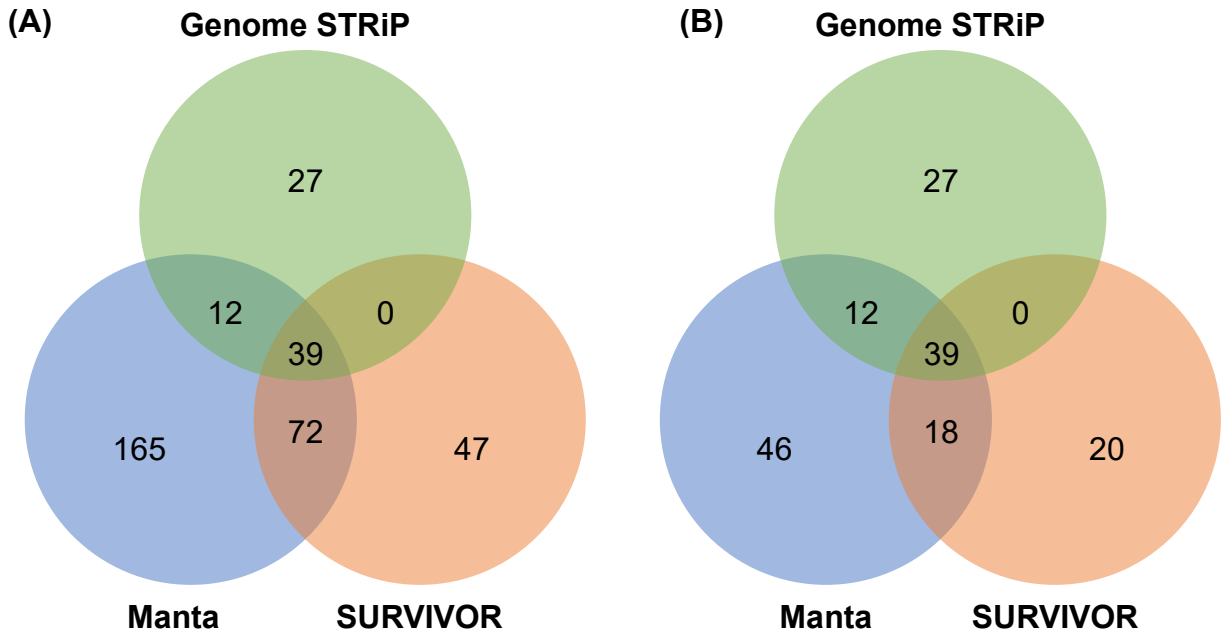


Figure 2.8: Overlap between the CNVs called by Genome STRiP (green), Manta (blue) and SURVIVOR (Delly, Lumpy and GATK) (orange) for (A) all CNVs (n = 362) and (B) CNVs between 1 kb and 100 kb in size (n = 162).

Each of the tools used have different abilities to detect different size ranges and types of CNVs. For example, Genome STRiP is unable to call CNVs that are smaller than 1 000 bp, while the other tools can call CNVs down to 50 bp (*Figure 2.6*). Additionally, Genome STRiP was unable to call CNVs greater than 100 kb in size, while the other tools were able to call several CNVs in this size range (*Figure 2.6*). The majority of the CNVs that were unique to the Manta dataset (n = 106; 64.2%), as well as 34% (n = 16) of the CNVs that were unique to the SURVIVOR dataset were small CNVs (< 1 000 bp)

that would not have been detected by Genome STRiP. Therefore, as the tools have different size ranges of CNVs that they are able to call, we further assessed the overlap between the tools for CNVs between 1 kb and 100 kb in size, as this is the size range that all tools could detect CNVs within (*Figure 2.8B)*. In this size range, a total of 162 CNVs were identified across all three datasets. The percentage of CNVs that were common across all three datasets increased to 24.1%. Furthermore, the number of CNVs unique to the Manta dataset decreased to 28.4%.

2.3.3 Merged CNV dataset

2.3.3.1 CNV landscape in ADME genes

The merged CNV dataset consisted of 362 ADME CNVs, the majority of which were deletions (n = 284; 78.5%). Duplications accounted for 19.1% of the CNVs (n = 69) and the remaining nine CNVs (2.5%) were the mCNVs (*Table 2.3*). A file containing the CNVs can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Supplementary%20Files/Merged_Short_Read_ADME_CNVs.xlsx).

Table 2.3: Count, length and frequency of ADME deletions, duplications and mCNVs in the merged dataset.

CNV category	Count	Average length (kb)	AF < 1%	1% ≤ AF ≤ 5%	AF > 5%
Deletions	284 (78.5%)	343.8	147 (51.8%)	48 (16.9%)	89 (31.3%)
Duplications	69 (19.1%)	472.2	43 (62.3%)	15 (21.7%)	11 (15.9%)
mCNVs	9 (2.5%)	4.3	2 (22.2%)	4 (44.4%)	3 (33.3%)
	362	**359.9**	**192 (53%)**	**67 (18.5%)**	**103 (28.5%)**

Approximately half of the ADME CNVs in the merged dataset were rare, with an allele frequency of less than 1% (*Table 2.3*). This was followed by common CNVs (AF > 5%) and low frequency CNVs (1% ≤ AF ≤ 5%) which accounted for 28.5% (n = 103) and 18.5% (n = 67) of ADME CNVs respectively (*Table 2.3*). The proportion of rare duplications (62.3%) was higher compared to rare deletions (51.8%), and the converse is true for common variants, with the proportion of common deletions (31.3%) being almost double the proportion of common duplications (15.9%) (*Table 2.3*). The mCNVs that were identified ranged in allele frequency from 0.9% to 22.7%.

Additionally, 7.2% (n = 26) of the CNVs were observed at very high allele frequencies (>70%) in the studied populations. These CNVs were all small with an average length of only 324 bp, and all but one being less than 1 000 bp in size. Sixteen ADME genes were affected by these high-frequency CNVs, all of which were extended ADME genes with the exception of *UGT2B7* which is classified as a core ADME gene. This CNV affected the upstream region of *UGT2B7* and was a 324 bp deletion with an allele frequency of 79.5%.

A total of 169 ADME genes were affected by CNVs, including 25 (14.8%) core ADME genes and 144 (85.2%) extended ADME genes. The core ADME genes affected by CNVs included (1) Phase I metaboliser genes *CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C19, CYP2C8, CYP2C9, CYP2E1, CYP3A4, CYP3A5* and *DPYD*; (2) Phase II metabolised genes *GSTM1, GSTP1, NAT1, SULT1A1, TPMT, UGT1A1* and *UGT2B7*; and (3) transporter genes *ABCC2, ABCG2, SLC15A2, SLC22A2, SLC22A6, SLCO1B1* and *SLCO1B3*. Phase I metaboliser genes were most commonly affected (n = 71; 42%), followed by transporter genes (n = 45; 27%), Phase II metaboliser genes (n = 41; 24%) and lastly, modifier genes (n = 12; 7%) (*Figure 2.9*).

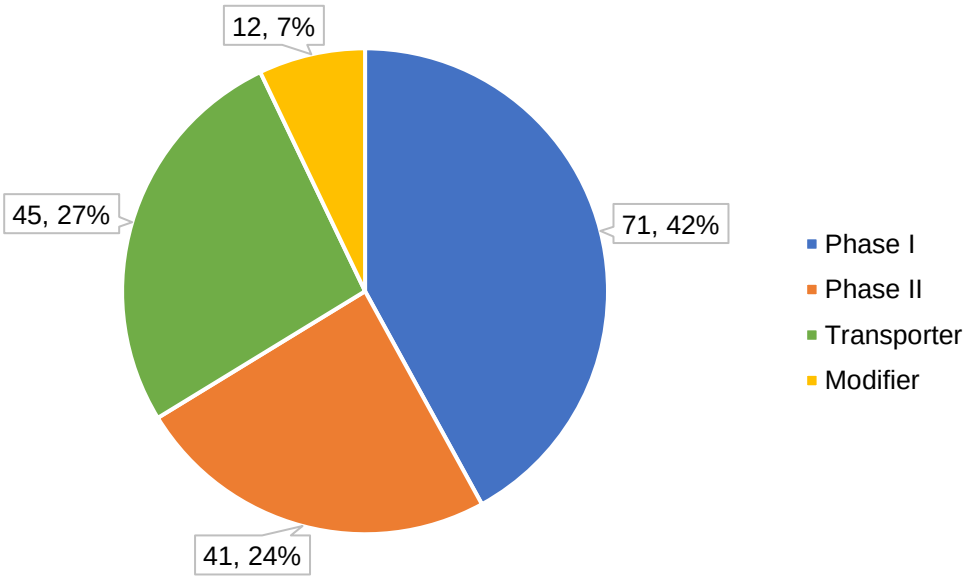


Figure 2.9: Breakdown of the type of ADME gene affected by CNVs (Count; Percentage).

The majority of ADME genes were not affected by any CNVs (n = 130; 43.5%). This was followed by ADME genes affected by only one CNV (n = 66; 22.1%), two CNVs (n = 41; 13.7%) and three CNVs (n = 21; 7%) (*Figure 2.10*). The remaining genes were affected by between four and 16 CNVs, with 1.7% (n = 5) of ADME genes being affected by more than 10 CNVs (*Figure 2.10*). A breakdown of the number of CNVs that each ADME gene was affected by can be seen in *Figure 2.11* below.

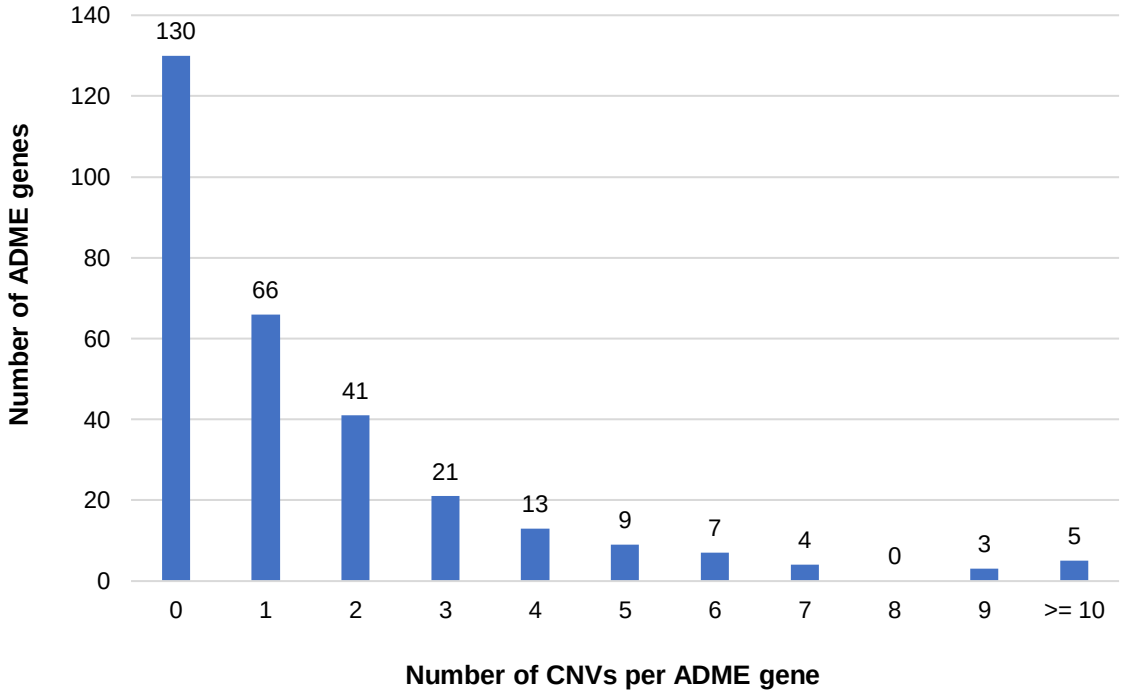


Figure 2.10: Number of CNVs per ADME gene.

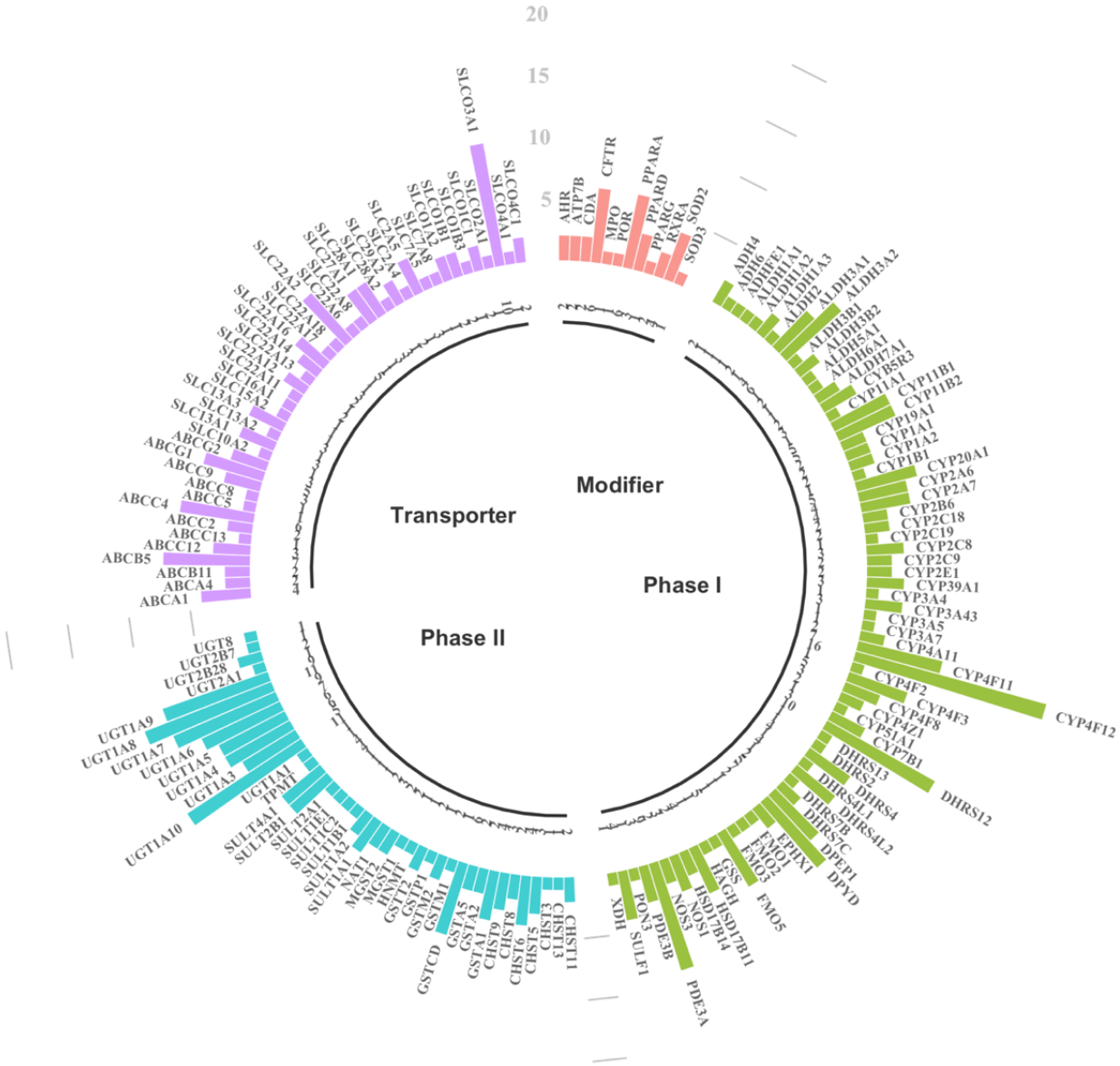


Figure 2.11: Number of CNVs per ADME gene for Phase I metaboliser genes (n = 71) (green), Phase II metaboliser genes (n = 45) (blue), transporter genes (n = 41) (purple) and modifier genes (n = 12) (red).

2.3.3.2 Regional analysis

The number of ADME CNVs per individual ranged from 32 to 68, with each individual harbouring an average of 48 ADME CNVs. The majority of individuals (n = 565; 59.3%) had between 40 and 50 CNVs affecting their ADME genes, and a further 30% (n = 286) of individuals harboured between 50 and 60 ADME CNVs (*Figure 2.12*). When separated by region, the average number of CNVs per individual was similar across SSA, with individuals from west Africa harbouring an average of 47 ADME CNVs, individuals from central and east Africa both harbouring an average of 48 ADME CNVs and individuals from southern Africa harbouring an average of 51 ADME CNVs.

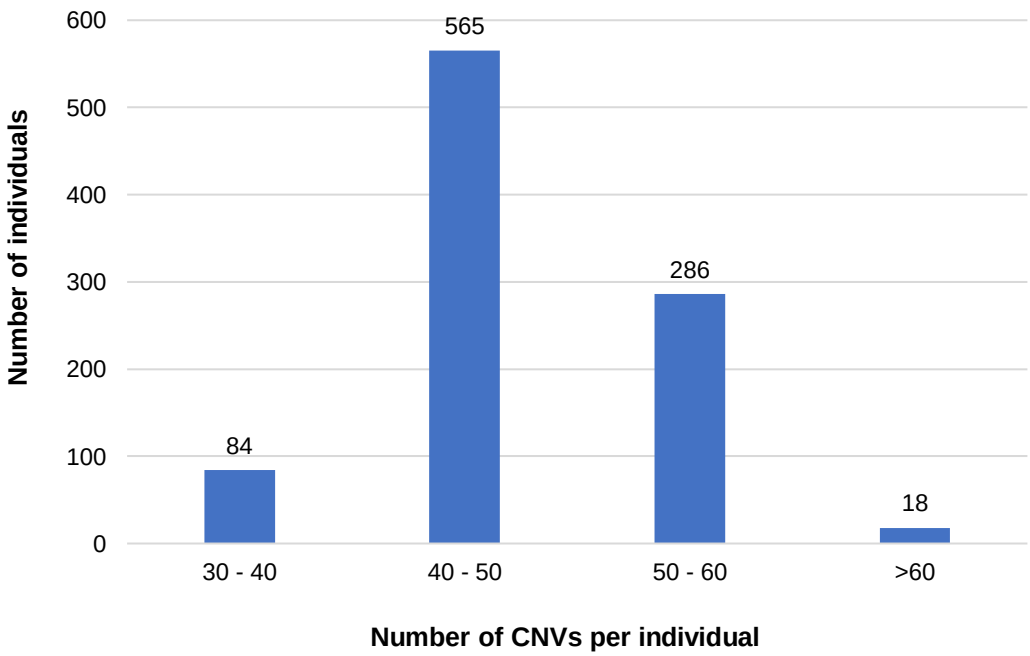


Figure 2.12: Number of ADME CNVs per individual.

The number of CNVs identified in each region of SSA can be seen in *Figure 2.13* below. It is important to note that the number of samples from each region is not equal. The highest number of samples in the dataset were of west African origin (n = 564) and west Africa harboured the highest number of CNVs (n = 310). East and southern African samples had a similar number of CNVs identified; however, there were 63 more samples from southern Africa compared to east Africa. Lastly, although there were only 30 central African samples, these samples had more than half the number of CNVs compared to west Africa, which had almost 20 times the number of samples. When scaled by the number of samples, west Africa had 0.6 CNVs per sample, east Africa had 1.6 CNVs per sample, southern Africa had 1.1 CNVs per sample and central Africa had 5.8 CNVs per sample.

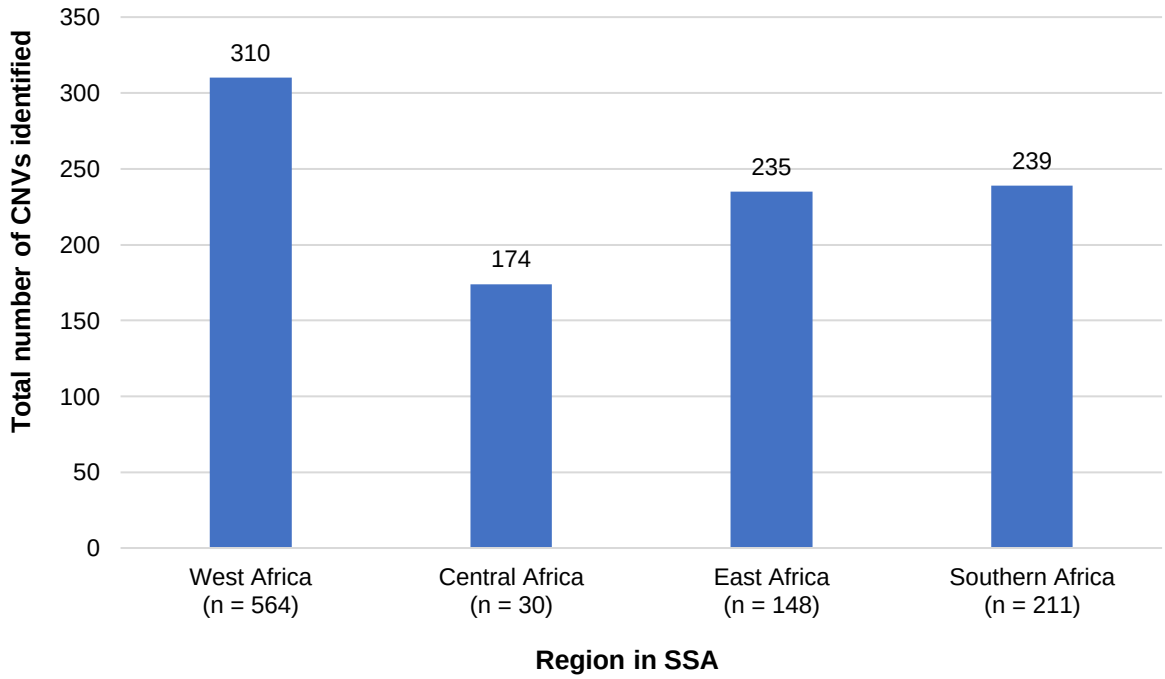


Figure 2.13: Total number of ADME CNVs identified in each region. The number of samples from each region is indicated in brackets below the region name. Note that the number of samples from each region are not equal.

A Venn diagram showing the overlap of CNVs identified in west, central, east and southern Africa can be seen in *Figure 2.14* below. A total of 34% of CNVs (n = 123) occurred in only a single SSA region, with west Africa having the highest number of unique CNVs (n = 74; 20.4%), followed by southern Africa (n = 21; 5.8%); east Africa (n = 20; 5.5%) and central Africa (n = 8; 2.2%). Just under half of the CNVs (n = 152; 42%) were common across all SSA regions. A total of 44 CNVs (12.1%) were common across three of the four SSA region with central, south and west Africa sharing the highest number of CNVs (n = 36; 9.9%). Lastly, 11.9% of CNVs (n = 43) were common across only two of the four SSA regions, with west and east Africa sharing the highest number of CNVs (n = 15; 4.1%), closely followed by west and southern Africa (n = 14; 3.9%).

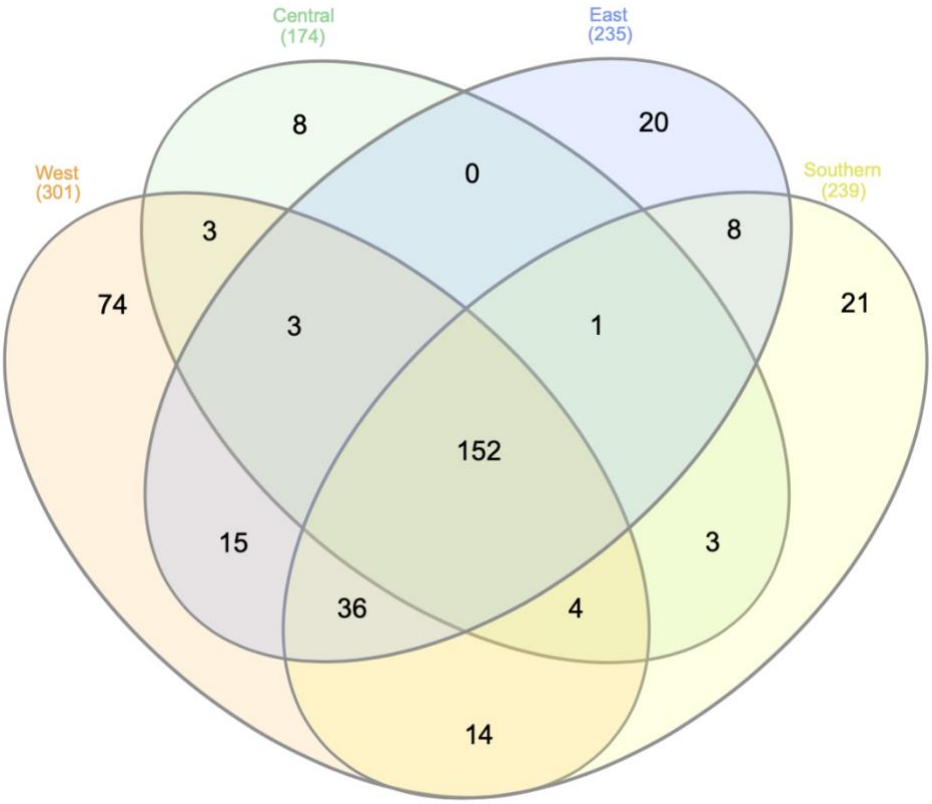


Figure 2.14: Overlap between ADME CNVs identified in individuals from west (orange), central (green), east (blue) and southern (yellow) Africa.

The CNVs that were common across all four regions (n = 152) affected 99 different ADME genes, including core ADME genes *ABCC2, ABCG2, CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2E1, DPYD, GSTM1, SLC22A2, SULT1A1* and *UGT2B7*.

2.3.3.3 Functional Annotation and Biological Implications

Size of ADME CNVs

The size of the CNVs in the merged dataset was assessed as this can have implications on the functional impact. The smallest CNV identified was 51 bp and the largest was 9.7 Mb in length. Duplications were larger than deletions, with an average length of 472.2 kb, while deletions had an average length of 343.8 kb (*Table 2.3)*. The mCNVs identified were far smaller, with an average length of only 4.3 kb (*Table 2.3*). The size range of the ADME CNVs in the merged dataset can be seen in *Figure 2.15* below. When the entire call set was assessed (*Figure 2.15A*), it can be seen that CNVs were most commonly (n = 219; 60.5%) between 101 and 10 000 bp in size. A further 15.2% (n = 55) of CNVs were between 10 001 and 100 000 bp and 9.9% (n = 36) of CNVs were greater than 100 000 bp in size (*Figure 2.15A*). Lastly, 14.4% (n = 52) of CNVs were small CNVs between 50 and 100 bp (*Figure 2.15A*). CNVs were further split into smaller size ranges including 50–1 000 bp (*Figure 2.15B*), 1 001–10 000 bp (*Figure 2.15C*) and 10 000–100 000 bp (*Figure 2.15D*). For small CNVs (<1000 bp), the most common size range was 50–100 bp (n = 52; 31.7%), followed by 101–200 bp (n = 28; 17.1%) (*Figure 2.15B*). For larger CNVs between 1 001 and 10 000 bp the most common size range was 1 001–2 000 bp (n = 33; 30.8%) followed by 2 001–3 000 bp (n = 21; 19.6%) and 4 001–5 000 bp (n = 15; 14%) (*Figure 2.15C*). Lastly, for large CNVs between 10 001bp–100 000 bp, CNVs were most commonly between 40 001–50 000 bp (n = 15; 27.3%), followed by 10 001–20 000 bp (n = 14; 25.5%) and 30 001–40 000 bp (n = 12; 21.8%) (*Figure 2.15D*).

Size versus allele frequency was plotted for all ADME CNVs in the merged dataset and it can be seen that there are a handful of extremely large CNVs (n = 23) greater than 1 Mb in length (*Figure 2.16A*). Most of these CNVs were observed at low allele frequencies; however, there were some that were observed at higher frequencies, up to 46.6% (*Figure 2.16A*). When CNVs greater than 100 000 bp were excluded (*Figure 2.16B*), the majority of CNVs greater than 10 000 bp were observed at low allele frequencies, but there is an increase in allele frequency for several CNVs around 40 000 bp. Lastly, when CNVs smaller than 10 000 bp were assessed (*Figure 2.16C*), small CNVs (<1000bp) had a vast range of allele frequencies. The allele frequencies decrease as the size increases; however, there are several CNVs around the 2 000 bp and 5 000 bp size range with increased allele frequencies (*Figure 2.16C*).

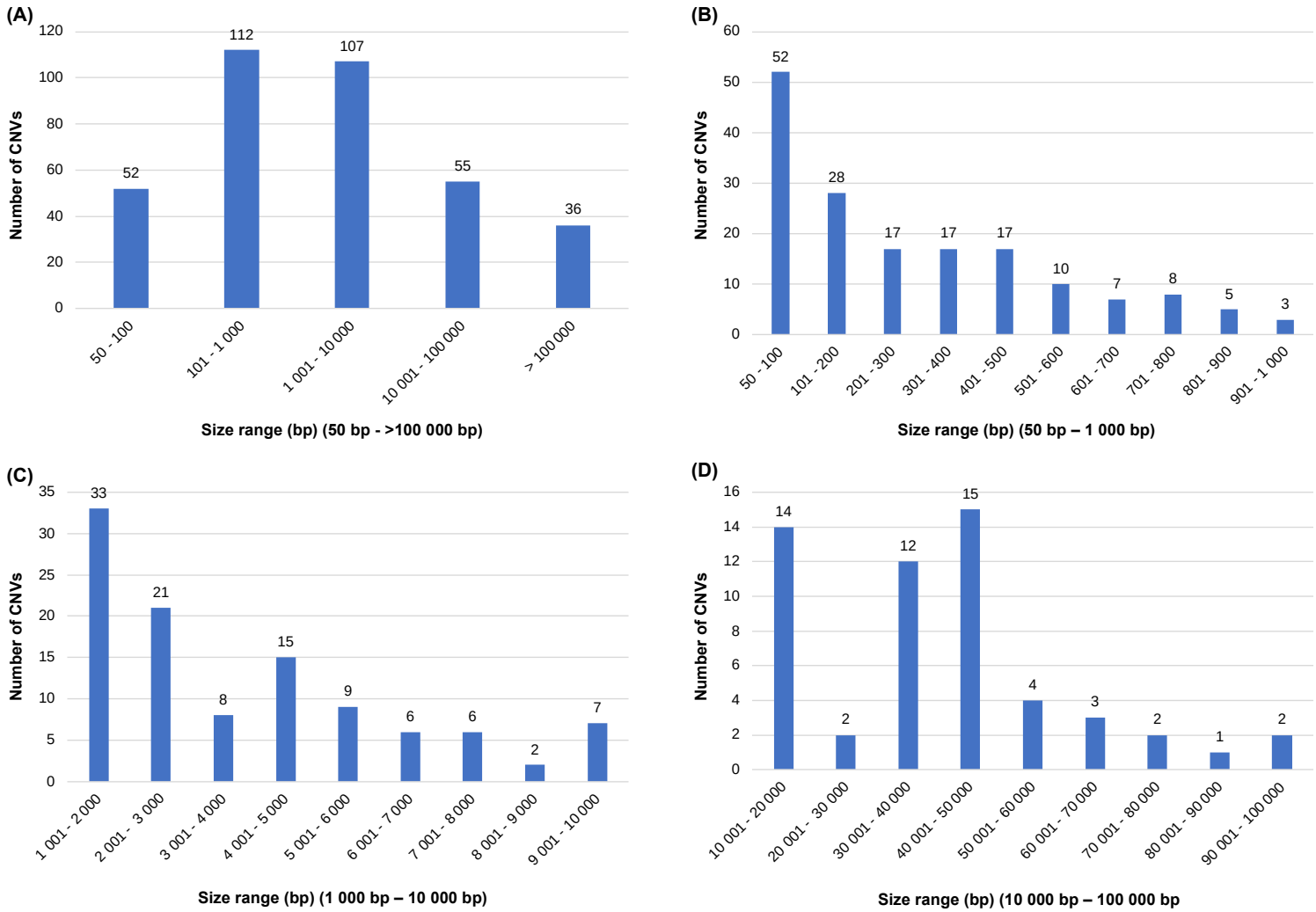


Figure 2.15: Size range of ADME CNVs (A) All CNVs; (B) CNVs between 50 and 1 000 bp in size; (C) CNVs between 1 000 and 10 000 bp in size and (D) CNVs between 10 000 bp and 100 000 bp in size.

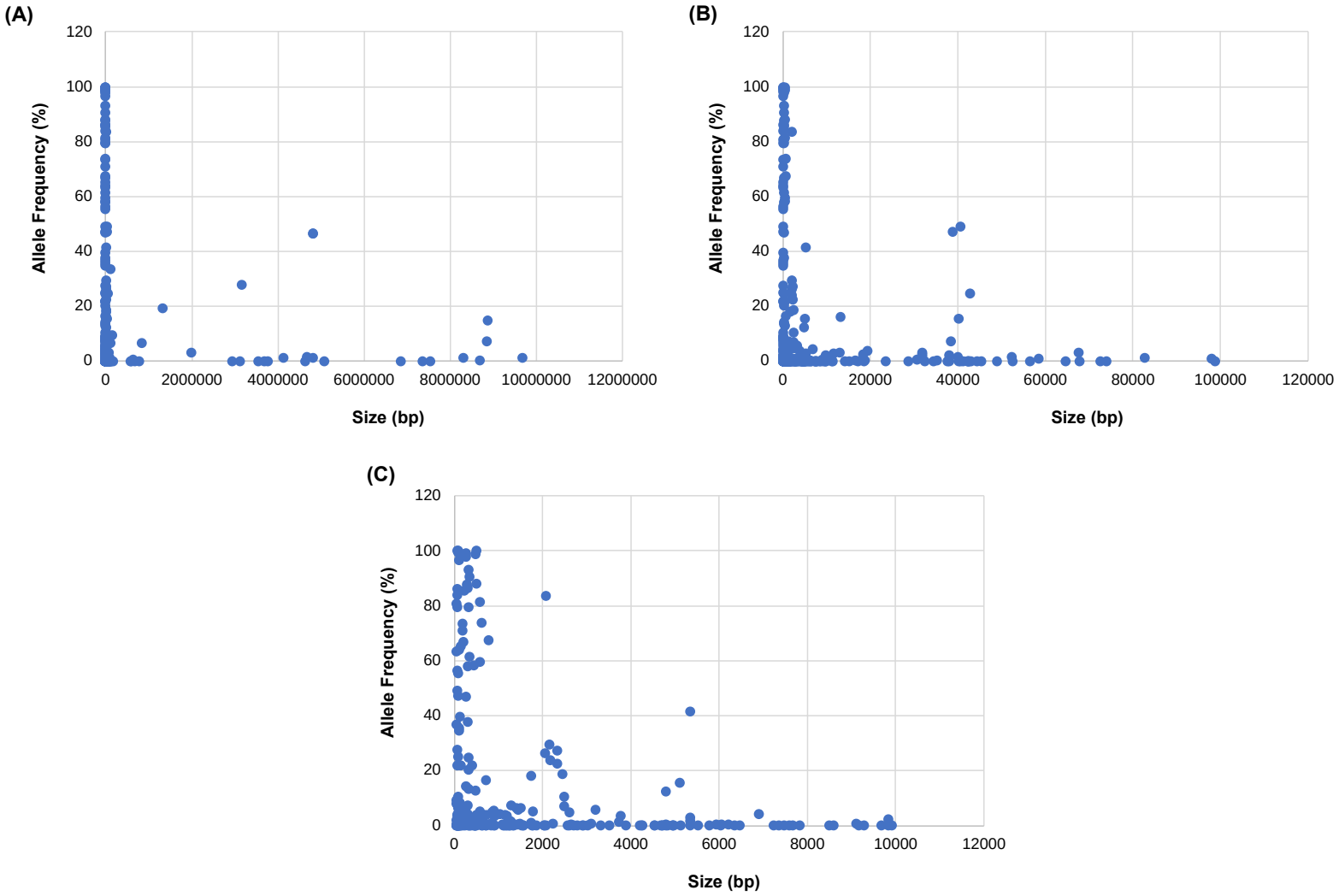


Figure 2.16: Size versus allele frequency for (A) all ADME CNVs; (B) ADME CNVs > 100 000 bp excluded and (C) ADME CNVs > 10 000 bp excluded.

Consequences of ADME CNVs

The most severe consequence of each CNV was obtained from Ensembl's VEP and can be seen in *Figure 2.17* below. Over half of the CNVs (n = 215; 59.4%) were intronic variants. This was followed by CNVs affecting the coding region of a gene/s (n = 53; 14.6%). Transcript ablation, or the complete loss of an entire gene/s, was the most severe consequence for 33 (9.1%) CNVs, and transcript amplification, or the complete gain of an entire gene/s, was the most severe consequence for 21 (5.8%) of the CNVs. A total of 57 ADME genes were either fully deleted (transcript ablations) or fully duplicated (transcript amplifications), nine of which were core ADME genes (*CYP1A1, CYP1A2, CYP2A6, CYP2C8, CYP2C9, CYP2E1, CYP3A4, GSTM1, GSTP1* and *NAT1*). Upstream and downstream variants accounted for 3.3% (n = 12) and 3% (n = 11) of CNVs. Stop losses (loss of a stop codon) were the most severe consequence for 2.8% (n = 10) of CNVs, and 3' and 5' untranslated region (UTR) variants accounted for 1.4% (n = 5) and 0.6% (n = 2) of CNVs.

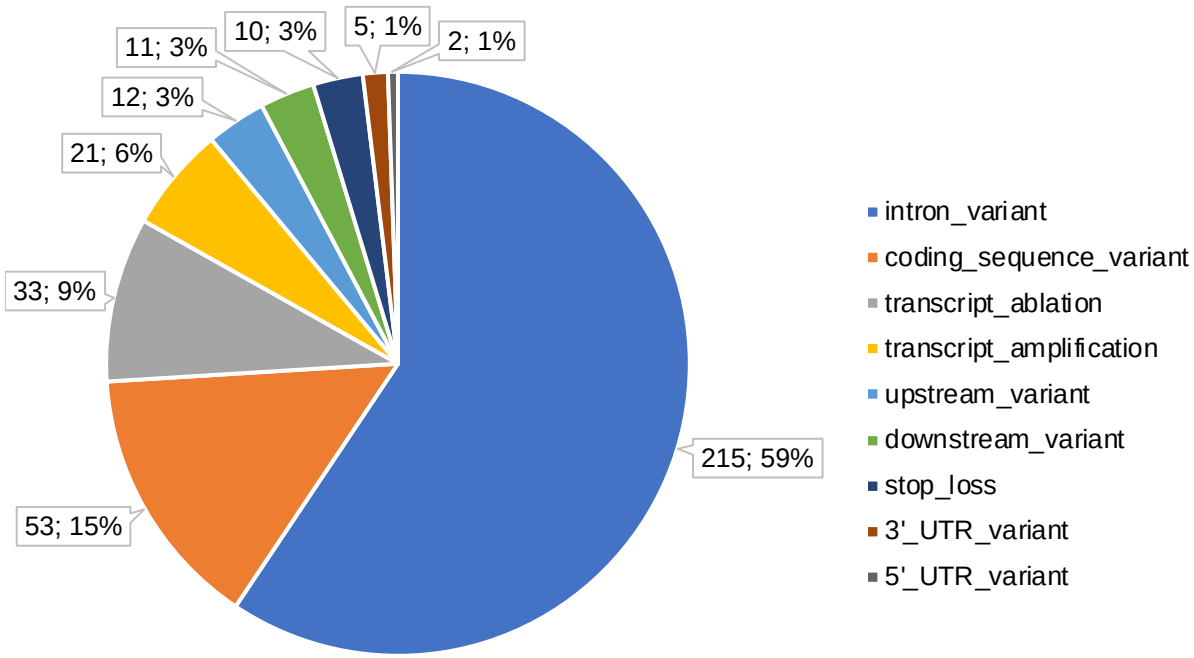


Figure 2.17: Most severe consequence of each CNV obtained from Ensembl's VEP (Count; Percentage).

Of the 152 CNVs that were found to be common across all SSA regions, 35 CNVs (23%) had severe consequences including affecting the coding region of a gene, transcript amplification, transcript ablation and the loss of the stop codon (stop loss).

Of the CNVRs affecting at least one exon, more CNVs affected terminal exons (i.e. affecting at least the 5' or 3' terminal exon) as opposed to internal exons only (*Figure 2.18*). Furthermore, of the CNVs affecting terminal exons, more CNVs affected the 5' terminal exons compared to the 3' terminal exons (*Figure 2.18*).

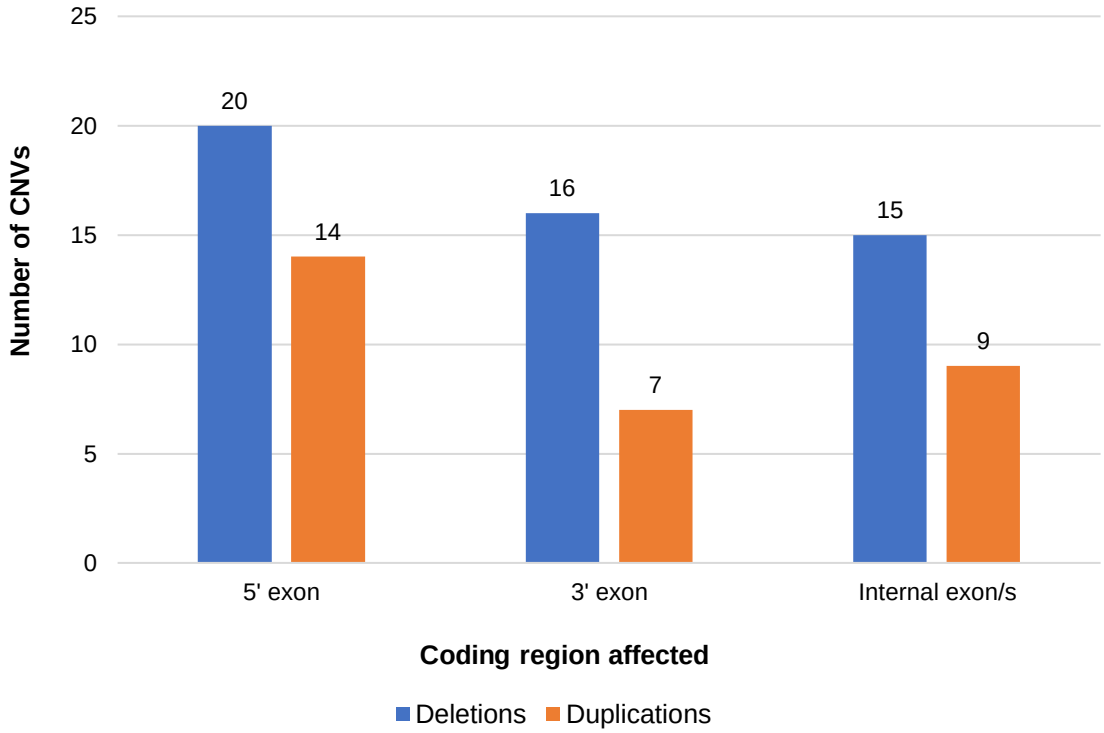


Figure 2.18: Breakdown of coding region affected by CNVs involving at least one exon.

Table 2.4 below shows which genes are involved in the ADME of drugs (antiretrovirals, antimalarials, anti-TB drugs, anti-cancer drugs, lipid-lowering drugs and anti-coagulants) that are frequently used to treat common communicable and non-communicable disease in Africa, as well as a breakdown of the CNVs that have been found within these genes. Important ADME genes to consider in the African context include the following CYP450 genes, *CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP3A4, CYP3A5* and *CYP2E1* as well as *GSTM1, UGT1A9, SULT1A1* and *SLCO1B1*. Numerous deletions and duplications were identified within these genes, some with severe functional consequences such as transcript amplification, transcript ablation and alterations to the coding regions of these genes. Additionally, several of the CNVs were identified at high frequencies in the studied SSA populations.

Table 2.4: Drugs used to treat common diseases in Africa, genes that are involved in the ADME of these drugs as well as the CNVs identified in these genes (type, size, frequency and gene region affected by the CNV).

ADME genes	Drugs																						CNV	Type	Size (bp)	Freq (%)	Gene region affected
	Antiretroviral						Antimalarial							Anti-TB		Anti-cancer				Lipid-lowering			Anticoaulant				
	Efavirenz	Nevirapine	Dolutegravir	Raltegravir	Ritonavir	Lopinavir	Chloroquine	Amodiaquine	Mefloquine	Quinine	Artemether	Lumefantrine	Artesunate	Isoniazid	Rifampicin	Paclitaxel	Docetaxel	Tamoxifen	Cisplatin	Oxaliplatin	Fluvastatin	Rosuvastatin	Atorvastatin	Warfarin				
CYP1A1	○																							chr15_71577734_80433233	Del	8855499	7.3	Transcript ablation
																								chr15_74721603_74745146	Del	23543	0.2	Exons 1-3/7; introns 1-2/6; 5' UTR
CYP1A2							○		●			○													chr15_71577734_80433233	Del	8855499	7.3	Transcript ablation
CYP2A6	○						○		○		●		●												chr19_40843626_40875428	Del	31802	2.2	Transctipt ablation
																								chr19_40848777_40880642	Del	31865	3.3	Exons 1-2/9; introns 1-2/8
																								chr19_40838483_40869134	Dup	30651	0.8	Transcript amplification
																								chr19_40848650_40880515	Dup	31865	2.5	Exons 1-3/9; introns 1-2/8
CYP2B6	●	●					○		○	○	○	○	○		○									chr19_40940696_41008574	Del	67878	0.3	Exons 1-4/9, introns 1-4/8; 5' UTR
																								chr19_40948220_41015852	Del	67632	3.2	Exons 1-8/9, introns 1-8/8; 5' UTR
CYP2C8								●							○	●								chr10_94930837_95095275	Dup	164438	0.05	Transcript amplification
																								chr10_95037827_95037901	Del	74	0.9	Intron 8/8
																								chr10_94954059_94954110	Del	89	10.5	Intron 5/8
CYP2C9	○						●		○			○			○			●			●	●		●	chr10_94930837_95095275	Dup	164438	0.05	Trasncript amplification
																								chr10_94954059_94954110	Del	51	0.2	Intron 5/8
CYP2C19							○		○			○						○							chr10_94816714_94825877	Del	9163	0.05	Exon 6/9; introns 5-6/9
CYP3A4	○	●			●	●	●	○	●	●	●	●	●		○	●	●	●					●		chr7_99738730_99787675	Del	48945	0.3	Transcript ablation
CYP3A5	○	○			●	●	●			●	●	●	●		○										chr7_99648651_99657137	Del	8486	0.05	Exons 11-12/13; introns 10-12/12
CYP2E1														○											chr10_133428994_133573823	Dup	144829	9.6	Transcript amplification
																								chr10_133523672_133523734	Del	61	1	Intron 1/10
GSTM1		○												●	○				○	○					chr1_109676376_109694830	Del	18454	2.5	Transcript ablation
UGT1A9											○		○												chr2_233625549_233677835	Del	52286	1.7	Exon 1/5; intron 1/4; 5'UTR
																								chr2_233625826_233678276	Dup	52450	0.2	Exon 1/5; intron 1/4; 5'UTR
SULT1A1																		●						chr16_28597983_28610977	Del	12994	3.3	Transcript ablation
																								chr16_28598180_28611297	Dup	13117	16.2	Transcript amplification
SLCO1B1						○																		chr12_21112163_21156501	Del	44338	0.05	Exons 1-2/15; introns 1-2/14; 5' UTR

● Major contributor to the drug's metabolism; ○ Minor contributor to the drug's metabolism

2.3.3.4 Novel and known CNV loci

Of the 362 CNVs in the merged dataset, 120 CNVs (33.2%) did not overlap with CNVs in the DGV and were classified as novel. The majority of novel CNVs were deletions (n = 91; 75.8%), while duplications accounted for the 22.5% (n = 27) of novel CNVs and mCNVs accounted for only 1.7% (n = 2) of novel CNVs. Half of novel CNVs were small CNVs (<1000 bp) (n = 60; 50%). There were also several larger CNVs with 29.2% of novel CNVs (n = 35) being between 1 000–10 000 bp in size, and a further 20.8% of novel CNVs (n = 25) being larger than 10 000 bp in size. In terms of allele frequency, the majority of novel CNVs (n = 76; 63.3%) were rare with an allele frequency of less than 1%. However, 25% (n = 30) of novel CNVs were common (AF > 5%) in the studied populations. A total of 102 ADME genes were affected by at least one novel CNV, including core ADME genes *ABCG2, CYP1A1, CYP1A2, CYP2C8, CYP2E1, CYP3A5, DPYD, GSTP1, SLC22A2, SLCO1B1, SLCO1B3, TPMT* and *UGT1A1*.

2.4 Discussion

It is well known that an individual's genetic variability within pharmacogenetically relevant genes plays a major role in the variable response to drugs that is observed amongst patients (Sim, Kacevska and Ingelman-Sundberg, 2012). Genetic variation within ADME genes explains to a large extent variability in drug efficacy as well as the presence of ADRs (Johansson and Ingelman-Sundberg, 2009). SNVs within ADME genes have been widely studied and characterised; however, CNV is an understudied, yet important source of variability that needs to be characterised in order to obtain a more complete set of pharmacogenetically relevant variants. In addition, African populations are generally understudied (Martin *et al.*, 2018), with no exception of CNVs. Over recent years, the introduction of NGS has resulted in the development of over 100 different tools for calling SVs, including CNVs, from short-read sequencing data (Mahmoud *et al*., 2019). By leveraging this approach in combination with a novel, African dataset, we have been able to provide the first assessment of the CNV landscape within ADME genes in a diverse set of SSA populations.

2.4.1 Differences between CNV calling tools

While many SV calling algorithms have been developed, no single algorithm can accurately and sensitively detect all types and all sizes of SVs (Pabinger *et al.*, 2014). Therefore, many studies have made use of multiple algorithms that use different approaches to call SVs and then merged the outputs (Kosugi *et al.*, 2019). In this study we made use of Genome STRiP (RP and RD), Manta (RP, SR and AS), Delly (RP and SR), Lumpy (RP, SR and RD) and GATK (RD) (Kosugi *et al.*, 2019) to call CNVs in ADME genes. Manta, Delly and Lumpy were included as they have been shown to have high levels of precision and recall (Kosugi *et al.*, 2019). Genome STRiP was included in this study as it calls CNVs across multiple samples, which is a different approach compared to all of the other tools. By using several different tools that make use of a combination of several different approaches, we aimed to obtain a more comprehensive dataset.

It has been shown that the CNVs detected by different tools can be highly variable (Coutelier *et al.*, 2021), as was observed in this study. We found that the number and size range of CNVs called by each tool varied widely, and that the three datasets overlapped poorly. This finding was also observed by Coutelier and colleagues who

made use of four different CNV calling tools on WGS data (Coutelier *et al.*, 2021). Of the 362 CNVs in our merged dataset, only 10.8% of CNVs were common to all three datasets (Genome STRiP, Manta and SURVIVOR (Delly, Lumpy and GATK merged)), and a further 23.2% of CNVs were common to two of the three datasets. Therefore, the majority of CNVs (66%) were unique to a specific dataset, with Manta having the highest proportion of unique CNVs (45.6%). In a study that made use of 11 different CNV calling algorithms on 36 WGS, 74.4% of CNVs were called only by a single tool, and only 9.5% of CNVs were called by three of the tools (Gabrielaite *et al.*, 2021). The percentage of CNVs called by three tools is comparable to our study while the concordance between our tools is slightly higher with only 66% of CNVs unique to one tool, compared to 74.4% in their study. However, this may be due to the fact that they made use of more CNV calling tools, each of which will have their own unique call set.

However, in a study that assessed just under 80 SV calling tools, they found that while some tools, such as Manta, could call deletions and duplications across all size ranges, others showed limited performance in specific size ranges (Kosugi *et al.*, 2019). An example of such a tool is Genome STRiP which is less effective for small deletions and duplications (Kosugi *et al.*, 2019). While Manta is able to call CNVs with a minimum size of 50bp, Genome STRiP is only able to call CNVs down to 1000bp. A total of 64.2% of the CNVs that were unique to the Manta dataset were small CNVs (< 1000bp) and would therefore not have been called effectively by Genome STRiP. SV calling tools also differ in their ability to call different types of SVs. While all of the tools that were used in this study have the ability to call both deletions and duplications, Genome STRiP is the only tool used that calls CNVs using multiple samples and is therefore able to call mCNVs.

Although the CNVs that were detected by each caller were highly variable in this study, the advantages of each tool were leveraged in order to obtain a more comprehensive dataset of ADME CNVs (Coutelier *et al.*, 2021).

2.4.2 Landscape of CNVs in ADME genes

In this study we investigated ADME CNVs in various SSA populations using several different tools. The results from the five tools were combined to produce a merged dataset which consisted of 362 ADME CNVs, the majority of which were deletions

(78.5%). The majority of individuals (89.3%) harboured between 40 and 60 CNVs within their ADME genes and, on average, individuals from southern Africa had the highest number of ADME CNVs (n = 51). In a study performed by Tremmel and colleagues that assessed CNV in ADME genes in 150 European individuals, each individual harboured an average of 4.1 CNVs in the 340 ADME genes that were assessed (Tremmel *et al.*, 2016). This is considerably lower than the number of ADME CNVs that were identified per individuals in our study, and can in part be explained by the fact that African individuals tend to harbour a greater level of genetic variation compared of European individuals (Campbell and Tishkoff, 2008). However, the Tremmel and colleagues study is not directly comparable to our study as they called CNVs from sequencing data generated using a targeted ADME-NGS panel (Tremmel *et al.*, 2016). The panel targeted the exons with exon/intron boundaries and selected 5′ gene regions of 340 ADME genes (Tremmel *et al.*, 2016), while our study called CNVs from WGS data, and thus included intronic and intergenic CNVs as well.

Just under 30% of the CNVs in the merged dataset were common, with an allele frequency of 5% or greater in the studied SSA populations. While a variety of common genetic variants that influence drug response have been previously reported, these do not fully explain the variability in drug response that is observed (Ingelman-Sundberg *et al.*, 2018). A recent study concluded that rare genetic variants are likely to account for a substantial proportion of the unexplained inter-individual variability that is observed in drug response (Ingelman-Sundberg *et al.*, 2018). Over half (53%) of the CNVs identified in the merged dataset were rare (AF < 1%) in SSA populations and may account for some of the unexplained variability mentioned above. The use of high-coverage WGS data in this study allowed us to identify rare variants that may otherwise have been missed if low-coverage WGS data was utilised.

Additionally, while common CNVs are important to consider in terms of precision public health, rare CNVs are important to consider in terms of precision medicine at the individual-level. An interesting example of this is the case of *DYPD.* In this study we identified seven different CNVs within the *DYPD*, six of which were intronic variants. The remaining CNV affected the coding and intronic region of the gene and resulted in the deletion of exon 11/23 and introns 10-11/22 (chr1_97572305_97577146). This CNV was observed in a single individual from west Africa, who was heterozygous for the deletion.

Fluorouracil is a chemotherapy agent that is used in the treatment of multiple cancer types (Dean and Kane, 2016). The DPD enzyme catalyses the rate-limiting step in fluorouracil metabolism to generate inactive metabolites (Dean and Kane, 2016). It is known that individuals with at least one copy of non-functional DPYD (DPYD*2A) may not metabolise fluorouracil at normal rates (Dean and Kane, 2016). Prolonged clearance of the drug puts these individuals at risk of life-threatening fluorouracil toxicity including bone marrow suppression, neurotoxicity, and even death (Saif *et al.*, 2007; Dean and Kane, 2016). This is considered to be a true pharmacogenetic syndrome, as symptoms are only present when an individual is administered the drug (Saif *et al.*, 2007). DPYD*2A is the most common mutation that has been associated with a deficiency of the DPYD enzyme (Saif *et al.*, 2007). A G → A base change at the splice recognition sequence of intron 14 results in exon skipping and a 165bp deletion of exon 14 in the mRNA (Saif *et al.*, 2007). This results in the production of a truncated protein with almost no enzyme activity (Saif *et al.*, 2007). The CNV identified in this study involves the deletion of exon 11 of *DPYD*. This truncated protein may have reduced or no enzyme activity depending on the importance of exon 11 in the protein. If so, it may have similar effects to the DPYD*2A variant. While fluorouracil was not always accessible and commonly prescribed in Africa, as of recently, Pfizer has agreed to provide fluorouracil at or near cost price to various African countries (Dyer, 2017). As this drug will be prescribed more frequently in Africa, the CNV that we have identified warrants further investigation due to the severity of the ADRs that are observed when there is a DYPD enzyme deficiency.

A third (33.2%) of the CNVs in the merged dataset did not overlap with CNVs in the DGV and were classified as novel. A study performed on over 6 000 European individuals found that 81% of the CNVs called from the WGS data had been previously reported in the DGV, therefore 19% of their CNVs were novel (Png *et al.*, 2019). We found a higher percentage of novel CNVs from far fewer sequences. This may be due to the higher levels of genetic diversity that are typically seen in African populations (Campbell and Tishkoff, 2008) but may also be because African populations are understudied and are therefore not well represented in databases such as the DGV (Mulder *et al.*, 2020).

A total of 34% of CNVs in the merged dataset were region specific, with west Africa harbouring the highest number of region specific CNVs. This may be due to the fact

that we had far more west African samples compared to other regions. The identification of region specific CNVs indicates that CNV is not uniform across African regions, and that African populations cannot be treated as an homogenous entity. However, we did identify CNVs that were present across all regions, with 42% of the CNVs in the merged dataset occurring across west, central, east and southern Africa. Of these CNVs that were common to all regions, several core ADME genes were affected and a large proportion (23%) had severe predicted consequences including transcript amplification, transcript ablation, stop loss, and coding regions affected. As these CNVs are observed across African regions and have potential functional implications, further investigations into these CNVs is warranted.

Interestingly, central Africa harboured a large number of CNVs in relation to the number of samples from the region. Although there were only 30 central African samples, these samples had more than half the number of CNVs compared to West Africa, which had almost 20 times the number of samples. The majority of samples from central Africa were from Cameroon. The high number of CNVs identified in individuals from this region compared to other regions may be due to the movement of Bantu-speaking individuals from Cameroon into the rest of SSA approximately 5 000 years ago (Choudhury *et al.*, 2018).

In this study a large proportion of ADME genes (n = 169) were found to have CNVs which may be explained by the fact that, in general, ADME genes are more variable and less sensitive to negative selection (Hovelson *et al.*, 2017). A study performed by Santos and colleagues assessed the CNV landscape in 208 ADME genes (Santos *et al.*, 2017). They utilized CNVs that had previously been called from 2 504 WGSs and 59 898 whole exome sequences (WESs) from various populations (non-Finnish Europeans, Finns, Africans, South Asians, East Asians, and admixed Americans) and found that of the 208 ADME genes that were included in their study, 201 harboured CNVs (Santos *et al.*, 2017). In our study, fewer ADME genes were found to have CNVs (n = 169), with a larger proportion of ADME genes containing no CNVs. However, this may be due to the far higher number of samples assessed by Santos and colleagues (Santos *et al.*, 2017). CNV calling in a larger number of study participants may increase the number of CNVs identified in ADME genes.

Furthermore, their study showed that 47% of the CNVs assessed were deletions, while the remaining 53% were duplications (Santos *et al.*, 2017). While they detected slightly more duplications compared to deletions, the majority of CNVs identified in this current study were deletions, which accounted for 78.5% of CNVs in the merged dataset.

2.4.3 Functional annotation and biological implications of CNVRs

The CNVs in the merged dataset ranged in size from 51 bp to 9.7 Mb, and just under half of the CNVs were small (<1 000 bp). We observed increased frequencies of several CNVs around the 2 kb and 5–6 kb size ranges as well as around the 40 kb size range. Similar patterns in CNV frequency were observed by Gonzaga-Jauregui and colleagues, specifically, between 10–30 kb, instead of 40 kb as in this current study (Gonzaga-Jauregui, Lupski and Gibbs, 2012). Additionally, duplications were larger than deletions, which was also observed by Santos and colleagues (Santos *et al.*, 2017). It is generally accepted that the larger the CNV, the greater the clinical significance; however, there are cases where large CNVs are clinically benign and small CNVs have significant function implications (Nowakowska, 2017).

CNVs can have an effect on phenotype by directly affecting gene dosage, thus altering the level of gene expression, and may also have significant impacts on protein sequence, structure and function (Zhao *et al.*, 2013; Nowakowska, 2017). The gene elements affected by CNVs were studied as this can influence the functional impact of the CNVs (Cabianca and Gabellini, 2010b). If an entire gene(s) locus is involved, the gene dosage may be altered resulting in either gene amplification or deletion (Stranger *et al.*, 2007; De Smith *et al.*, 2009). In this current study we identified 54 CNVs (14.9%) that affected an entire ADME gene/s. Full gene deletions are almost invariably damaging (Saif *et al.*, 2019), and the core ADME genes that were fully deleted included *CYP1A1, CYP1A2, CYP3A4, GSTM1* and *GSTP1.* While full gene deletions are almost always damaging, this is not the case for full gene duplications (Saif *et al.*, 2019). However, even though they are not always considered deleterious, they can have important implications in terms of pharmacogenetics. For example, when an individual has more than two functional copies of an ADME gene, they are considered to be an ultra-rapid metabolizer (UM) (Jarvis, Peter and Shaman, 2019). For drugs administered in their active form, UMs may require an increased dose as a result of the increased levels of metabolism of the active drug into its inactive metabolites

(Jarvis, Peter and Shaman, 2019). Opposingly, for prodrugs, lowered dose may be required for UMs, as the increased metabolism of the prodrug into its active form may result in toxic levels of the drug in the body (Jarvis, Peter and Shaman, 2019). The core ADME genes that were fully duplicated included *CYP2A6, CYP2C8, CYP2C9, CYP2E1 and NAT1*.

However, the vast majority of CNVs did not affect an entire gene, but rather affected only a portion of a gene. Of the partial-gene CNVs, intronic CNVs (CNVs not affecting any exons) were the most abundant form of CNV within the ADME genes. As variants within introns are generally less damaging than those affecting coding regions, the enrichment of intronic CNVs may be explained as a consequence of negative selection of CNVs in exonic regions, whereby deleterious variants are selectively removed from the coding regions of genes (Rigau *et al.*, 2019). The apparent enrichment of intronic CNVs may also be explained by the fact that introns are typically longer than exons, with an average size of 3 kb and 100–200 bp respectively (Lim and Burge, 2001). Intronic variants may not be as benign as once thought. It is now well recognized that introns contain regulatory elements that may play a role in the control of gene expression and thus CNVs involving an intron/s may result in changes to gene expression (Rigau *et al.*, 2019). Furthermore, it has also been shown that CNVs within introns can lead to altered splicing or an alteration in the production of non-coding RNA (ncRNA) (Stranger *et al.*, 2007; De Smith *et al.*, 2009; Cabianca and Gabellini, 2010b).

While intronic CNVRs were the most abundant, 15% of CNVs affected the coding region of an ADME gene/s and exons were more affected by deletions as compared to duplications. CNVs affecting exons can result in changes to protein structure and function, thus leading to the production of an aberrant protein (De Smith *et al.*, 2009; Cabianca and Gabellini, 2010b). Functional impact may be further influenced by which exons are affected. More CNVs (both deletions and duplications) affected the terminal exons of the genes compared to only internal exons. CNVs involving only internal exons are likely to disrupt the reading frame and are thus likely to be damaging (Saif *et al.*, 2019). Furthermore, more CNVs affected the 5’ terminal exons as opposed to the 3’ terminal exons. Functionally, deletions involving the 5’ terminal exons of genes are almost always deleterious whereas duplications at the 3’ terminal exons may not be damaging, as the functional gene structure may remain intact (Saif *et al.*, 2019).

Calling CNVs from WGS data is just the first step. There is a great need for comprehensive databases and accurate tools for variant annotation (Coutelier *et al.*, 2021). While various tools exist for predicting the pathogenicity of SNVs, constructing similar tools for SVs is difficult due to the diversity of variant size and type. Furthermore, precise breakpoint detection of SV remains difficult with the current sequencing technology, resulting in uncertainty about their exact location. There have been some attempts at SV impact prediction. For example, VEP performs superficial consequence prediction for insertions, deletions and duplications. However, to date, no existing method provides a quantitative pathogenicity score for SV, making the prediction of functional impact difficult.

2.4.4 Important ADME CNVs in an African context

The African continent is severely affected by disease – both communicable and non-communicable (Alessandrini and Pepper, 2014). The rates of infectious diseases such as TB, HIV/AIDS and malaria remain high (Avert, 2020; World Health Organization, 2020b, 2020a), and the rates of non-communicable diseases (NCDs) such as cardiovascular diseases, diabetes mellitus, respiratory diseases, cancer and neuropsychiatric disorders are rapidly rising (Mayosi *et al.*, 2009; Nyirenda, 2016). Due to the high prevalence of these diseases in Africa, genetic variation within genes that are involved in the ADME of the drugs used to treat these common diseases is important to consider. The important drugs in an African context and the genes involved in their ADME, as well as the CNVs identified within these genes are summarized in *Table 2.4*. In the following section we will discuss the importance of each gene, which has been studied extensively by Bains (2013), Dandara and colleagues (2014) and Elliot, Mahungu and Owen (2017), as well as the CNVs affecting each gene.

CYP450 genes

CYP1A1 is a minor contributor to the metabolism of efavirenz, an ARV drug, while *CYP1A2* is involved in the metabolism of lumefantrine (Abdullahi *et al.*, 2020) and chloroquine and is a primary metaboliser of mefloquine, all of which are anti-malarial drugs. In this study we identified a deletion (chr15_74721603_74745146) that affected exons 1–3/7, introns 1–2/6 and the 5’ UTR of *CYP1A1* which was observed at an allele frequency of 0.2%. Deletions involving the coding regions of these genes have been

reported at frequencies of 0.03% in African individuals (Santos *et al.*, 2017). We also identified a large deletion (chr15_71577734_80433233), that spanned both *CYP1A1* and *CYP1A2,* which was common (AF = 7.3%) across the studied populations and was observed in individuals from west, central, east and southern Africa. Severe loss of function variants are not commonly observed in these genes (Ugartondo *et al.*, 2021), perhaps due to their important role in the deactivation of most endogenous compounds as well as xenobiotics (He and Feng, 2015). Therefore, the complete loss of these genes in a large proportion of the studied populations is an interesting finding that needs to be further investigated due to the potential functional impact and importance of these genes in the metabolism of several important drugs using in Africa.

CYP2A6 is a major contributor to the metabolism of anti-malarial drugs artemether and artesunate as well as a minor contributor to the metabolism of chloroquine and mefloquine (Navaratnam *et al.*, 2000; Ali *et al.*, 2010). It also plays a minor role in the metabolism of ARV drug, efavirenz. Both full gene deletions and duplications of the *CYP2A6* gene have been reported across various populations, with the full gene deletion being reported at an allele frequency of 3.8% in African American individuals (Martis *et al.*, 2013). In this study we identified various CNVs affecting this gene, the first of which was a deletion (chr19_40843626_40875428) that involved 98.7% of the gene (all exons, all introns, 5' and 3' UTRs) and was observed at a frequency of 2.2% in the studied populations. This is slightly lower than the allele frequency of the full gene deletion observed in African American individuals (Martis *et al.*, 2013). The full gene deletion of *CYP2A6* is known as *CYP2A6*4* and results in null enzyme activity and thus a poor metaboliser (PM) phenotype (Mittal *et al.*, 2015). This CNV is thus likely to affect the efficacy and toxicity of *CYP2A6* substrates such as important antimalarial drugs artemether and artesunate. We also identified a deletion (chr19_40848777_40880642) that affected only the terminal portion of the gene (exons 1-2/9 and introns 1-2/8), which was observed at an allele frequency of 3.3%. Full gene duplications of *CYP2A6* are less common and are generally observed at a frequency of < 2% across populations (Martis *et al.*, 2013). In this study the full gene duplication (chr19_40838483_40869134) of *CYP2A6* was observed at an allele frequency of 0.8%, which is in line with other studies (Martis *et al.*, 2013). Alleles that confer increased enzyme activity such as gene duplications result in increased rates of substrate metabolism. This has been extensively studied with respect to nicotine metabolism and smoking behaviours, with such alleles resulting in increased rates of

nicotine metabolism and thus increase cigarette consumption and depth of inhalation (Rao *et al.*, 2000). We also identified a duplication (chr19_40848650_40880515) affecting the terminal end of the gene (exons 1–3/9 and introns 1–2/8) which was observed at an allele frequency of 2.5%. All four CNVs identified in this study were observed in individuals across west, central, east and southern Africa.

CYP2B6 is key to the metabolism of various drugs, most predominantly efavirenz and nevirapine (Ciccacci *et al.*, 2017) which are both ARVs. It is also involved, although not predominantly, in the metabolism of anti-malarial drugs quinine, chloroquine, artemether, artesunate, lumefantrine and mefloquine (Navaratnam *et al.*, 2000; Ali *et al.*, 2010; Abdullahi *et al.*, 2020). Lastly, *CYP2B6* plays a role in the metabolism of rifampicin, an anti-TB drug (Bibi, 2008; Cho *et al.*, 2016). A recent study performed by Santos and colleagues found that in African individuals, deletions and duplications in *CYP2B6* were the most frequent, with frequencies of 0.9% and 0.4% respectively (Santos *et al.*, 2017). In this study, two CNVs within *CYP2B6* were identified, both of which were deletions (chr19_40940696_41008574 and chr19_40948220_41015852). One of the CNVs was a 67.6 kb deletion involving over 90% of the gene, including exons 1–8/9 and introns 1–8/8 as well as the 5' UTR (chr19_40948220_41015852). This CNV was observed at an allele frequency of 3.2% across the SSA populations, which is higher than the frequency observed by Santos and colleagues, and was observed in individuals across west, central, east and southern Africa. Efavirenz and nevirapine are two drugs commonly used a first-line treatments for HIV in LMICs such as those in SSA (Langmia *et al.*, 2021). Efavirenz has a narrow therapeutic window meaning that there is only a small difference between a therapeutic dose and a toxic dose (Langmia *et al.*, 2021). Therefore the risk of efavirenz-related ADRs including neurotoxicity, neuropsychiatric disorders and drug-induced lever disease, is high (Langmia *et al.*, 2021). Nevirapine is also associated with several ADRs and studies have recommended a reduced dose of the drug for individuals with PM variants in *CYP2B6* (Schipani *et al.*, 2011). Loss of function alleles, such as the gene deletion identified in this study, result in reduced metabolism and thus increased exposure to the drug and an increased risk of ARDs (Langmia *et al.*, 2021). Therefore, this CNV is extremely important to consider in an African context, where the frequency of the variant is relatively high and these drugs are commonly prescribed.

CYP2C8 is a primary metaboliser of amodiaquine, an anti-malarial drug, and paclitaxel, an anti-cancer drug (Spratlin and Sawyer, 2007), and it also plays a role in the metabolism of anti-TB drug rifampicin (Bibi, 2008; Cho *et al.*, 2016). In this study we identified two small intronic deletions in *CYP2C8*, the first was 74 bp in size, affected intron 8/8 and observed at an allele frequency of 0.9% (chr10_95037827_95037901), and the second was 89 bp in size, affected intron 5/8 and was observed at an allele frequency of 10.5% (chr10_95049857_95049947). The second CNV was common to all studied SSA regions. As mutations in introns do not affect the sequence of a protein itself, these variants are often ignored when assessing pathogenic variation (Rigau *et al.*, 2019). However, introns have been shown to play an important role in gene expression and studies have shown that intronic deletions can be associated with both increased or decreased expression of the affected gene (Rigau *et al.*, 2019). Using paclitaxel as an example, individuals who have reduced-activity variants in *CYP2C8* show slower clearance of the drug and are therefore at an increased risk of paclitaxel-induced ADRs (Marcath *et al.*, 2019). The same phenotype may be observed with reduced gene expression as a result of an intronic deletion. Therefore, the functional impact of these intronic deletions need to be further evaluated in order to assess whether these variants, particularly the high frequency variant, are important to consider in a pharmacogenetic context.

CYP2C9 is involved in the metabolism of several drugs and is considered as the most important gene in the metabolism of anticoagulant drug, warfarin (Takahashi and Echizen, 2003). It also plays a major role in the metabolism of fluvastatin and rosuvastatin, which are both lipid-lowering drugs (Scripture and Pieper, 2001; Olsson, McTaggart and Raza, 2002; Yan *et al.*, 2018). Furthermore, *CYP2C9* is primarily involved in the metabolism of anticancer drug, tamoxifen (Cronin-Fenton, Damkier and Lash, 2014). This gene is a major contributor to the metabolism of anti-malarial drug, chloroquine and also contributes in a minor way to the metabolism of efavirenz (ARV), mefloquine (anti-malaria drug) and lumefantrine (anti-malaria drug) (Abdullahi *et al.*, 2020). Lastly, *CYP2C9* plays a role in the metabolism of anti-TB drug rifampicin (Bibi, 2008; Cho *et al.*, 2016). In this study we identified a large duplication (chr10_94930837_95095275) that spanned *CYP2C8* and *CYP2C9* resulting in the transcript amplification of both genes. This duplication was observed at a frequency of 0.05%. While rare, this variant may have significant pharmacogenomic impacts for the individuals carrying it, as increased activity of these genes may require a higher dose

of the above-mentioned drugs in order to achieve therapeutic levels, or a reduced dose in the case of prodrugs. Additionally, a small, intronic CNV was identified in *CYP2C9.* This CNV was 51 bp in size, involved the deletion on intron 5/8 of the gene and was observed at an allele frequency of 0.2% (chr10_94954059_94954110).

CYP2C19 is a contributor to the metabolism of lumefantrine, chloroquine and mefloquine, all of which are anti-malarial drugs, as well as anti-cancer drug tamoxifen (Cronin-Fenton, Damkier and Lash, 2014). A deletion (chr10_94816714_94825877) of the coding and intronic region of this gene (exon 6/9 and introns 5-6/9) was identified at an allele frequency of 0.050%. Exonic deletions within *CYP2C19* have been reported at a similar allele frequency of 0.037% in DGV (Botton *et al.*, 2019).

CYP3A4 plays a role in the metabolism of several drugs used in the treatment of various common diseases in SSA. This gene plays a key role in the metabolism of nevirapine, ritonavir and lopinavir as well as a minor role in the metabolism of efavirenz (Ciccacci *et al.*, 2017) which are all used in the treatment of HIV/AIDS. Additionally, it is predominantly involved in the metabolism of tamoxifen (Cronin-Fenton, Damkier and Lash, 2014), paclitaxel (Spratlin and Sawyer, 2007) and docetaxel (Kenmotsu and Tanigawara, 2015) which are all anti-cancer drugs, as well as atorvastatin, a lipid-lowering drug (Korhonova, Doricakova and Dvorak, 2015; Yan *et al.*, 2018). *CYP3A4* is also predominantly responsible for the metabolism of anti-malarial drugs artemether, artesunate, lumefantrine, mefloquine, quinine and chloroquine and is a contributor to the metabolism of amodiaquine, also an anti-malarial drug (Navaratnam *et al.*, 2000; Ali *et al.*, 2010; Zhou *et al.*, 2019; Abdullahi *et al.*, 2020). Lastly, *CYP3A4* plays a role in the metabolism of anti-TB drug rifampicin (Bibi, 2008; Cho *et al.*, 2016). In this study we identified a full gene deletion (chr7_99738730_99787675) that spanned the entirety of *CYP3A4.* This heterozygous transcript ablation CNV was identified at an allele frequency of 0.3% and was only seen in individuals from west Africa. Deletions of this genes have not been previously reported, with the exception of the identification of an exonic deletion within the gene in a Colombian individual (AF = 0.4%) using MLPA (Ramírez *et al.*, 2019). This is an important CNV to consider in an African context due to the large number of important drugs that are metabolised by *CYP3A4* and the fact that a full gene deletion may have significant impacts on drug efficacy and toxicity.

CYP3A5 is a major contributor in the metabolism of ritonavir and lopinavir (Bains, 2013; Dandara *et al.*, 2014) and provides minor contributions to the metabolism of efavirenz and nevaripine (Ciccacci *et al.*, 2017), all of which are ARVs. In terms of the anti-malarial drugs, *CYP3A5* is a primary metaboliser for chloroquine, quinine, artemether, artesunate and lumefantrine (Navaratnam *et al.*, 2000; Ali *et al.*, 2010; Zhou *et al.*, 2019; Abdullahi *et al.*, 2020). *CYP3A5* also plays a role in the metabolism of anti-TB drug rifampicin (Bibi, 2008; Cho *et al.*, 2016). In this study we identified an 8 kb coding and intronic sequence deletion in *CYP3A5* (exons 11–12/13 and introns 10–12/12) with an allele frequency of 0.05% (chr7_99648651_99657137). Deletions within this gene were reported at a frequency of 0.02% in African populations by Santos and colleagues (Santos *et al.*, 2017).

Lastly, *CYP2E1* is involved in the metabolism of isoniazid, an anti-TB drug (Brito *et al.*, 2014; Wang *et al.*, 2016). In this study we identified a full gene duplication (chr10_133428994_133573823) at an allele frequency of 9.6%, as well as a small (61 bp), partial gene deletion involving intron 1/10 (chr10_133523672_133523734). Both of these CNVs were observed in individuals from west, central, east and southern Africa. The duplication of *CYP2E1* has been previously reported at a lower allele frequency of 1.7% in European patients (Tremmel *et al.*, 2016). A common ADR associated with isoniazid is isoniazid-induced hepatotoxicity (IIH) and mutations that have been found to enhance CYP2E1 enzyme activity have been linked to an increased risk of IIH. Therefore, a full gene deletion may have a similar impact, warranting further investigation of this CNV due to the high frequency at which it is observed across SSA and the common use of this drug.

GSTs

GSTM1 and *GSTT1* are minor contributors to the metabolism of nevaripine (ARV) and are major contributors to the metabolism isoniazid (anti-TB drug) (Brito *et al.*, 2014; Wang *et al.*, 2016). Although not the primary metabolisers, these genes can catalyse one of the reactions in the breakdown of cisplatin and oxaliplatin, both of which are anti-cancer drugs (Allocati *et al.*, 2018). The *GST*s also play a role in the metabolism of anti-TB drug rifampicin (Bibi, 2008; Cho *et al.*, 2016). Previous studies have shown that >27% of individuals are homozygous for *GSTM1* gene deletions and >37% of individuals are homozygous for *GSTT1* gene deletions (Garte *et al.*, 2001; Piacentini

et al., 2011). In this study CNVs in *GSTT1* were excluded as this gene is overlapped by an alternate contig. We did identify a full gene deletion (chr1_109676376_109694830) of *GSTM1* which was observed in individuals from west, central, east and southern Africa; however, the frequency was significantly lower than previously reported (2.5%). *GSTM1* gene duplications have been shown to be rare (He, Hoskins and McLeod, 2011), and no *GSTM1* duplications were identified.

UGTs

UGT1A9 and *UGT2B7* play a role in the metabolism of artemether and artesunate (anti-malarial drugs) (Ilett *et al.*, 2002; Zang *et al.*, 2014) and *UGT2B7* also contributes to the metabolism of efavirenz (ARV) (Bains, 2013; Dandara *et al.*, 2014; Elliot, Mahungu and Owen, 2017). Nine CNVs were identified in *UGT1A9*, seven of which were intronic CNVs. The remaining two CNVs included a deletion (chr2_233625549_233677835) and duplication (chr2_233625826_233678276) of a similar region of the gene (exon 1/5, intron 1/4 and the 5'UTR) which were observed at an allele frequency of 1.7% and 0.2% respectively. Exonic deletions and duplications in this gene have been reported at frequencies of 0.1% and 0.01% in African populations respectively (Santos *et al.*, 2017). No CNVs were identified in *UGT2B7*.

CNVs within *UGT2B28* have been previously reported in individuals of European descent at a frequency of 28.7% (Tremmel *et al.*, 2020). In this study we identified a full gene deletion of *UGT2B28* at an allele frequency of 33.8%. This CNV was observed in individuals from west, central, east and southern Africa. We also identified a partial gene deletion that affected 60.6% of the gene (exons 3-6/6, introns 2-5/5 as well as the 3' UTR) which was observed at an allele frequency of 6.8%.

Other genes

SULT1A1 is responsible for the conversion of tamoxifen metabolites into excretable forms (Cronin-Fenton, Damkier and Lash, 2014). Deletions of *SULT1A1* have been observed across multiple populations, with frequencies in African American populations being reported at 3.3% (Vijzelaar *et al.*, 2018). *SULT1A1* gene duplications are far more common and have been observed at a frequency of 60.9% in African American individuals (Gaedigk and Leeder, 2012; Vijzelaar *et al.*, 2018). In this study a deletion (chr16_28597983_28610977) affecting the majority of the gene (exons,

introns, 5' UTR and 3'UTR) was identified at an allele frequency of 3.3% in the studied populations. A duplication (chr16_28598180_28611297) of a very similar region was also identified and was observed at a frequency of 16.2%, which is lower than previously reported. Both the deletion and duplication were observed in individuals from west, central, east and southern Africa.

Lastly, most ARVs are transported by *ABCB1* and *ABCC2* and variations within these transporter genes can influence the safety and efficacy of the drugs they transport (Kis *et al.*, 2016). No CNVs were identified in *ABCB1,* while two intronic CNVs, a deletion (chr10_99814206_99814270) and duplication (chr10_99801858_99802084), were identified in *ABCC2*, both of which were observed in individuals from west, central, east and southern Africa.

Unfortunately, there are a number of important core ADME genes that are located in regions overlapped by alternate contigs. These include, but are not limited to, *CYP2D6*, *GSTT1, UGT2B15* and *UGT2B17*. Even though CNVs did occur within these genes, the calls could not be included with confidence due to quality control issues with the BAM files.

2.4.5 Conclusions

This work has illustrated the difficulties associated with accurately calling CNVs from short-read data, and how variable the results from different tools can be. The intrinsic limitations of calling CNVs from short-read sequencing data can only be overcome by other sequencing technologies or calling approaches (Coutelier *et al.*, 2021). In the following chapter, a long-read sequencing approach has been used in order to validate the CNVs identified in this chapter, gain a deeper understanding of the CNV landscape and determine whether long-read sequencing is required for accurate CNV calling.

Nevertheless, this study has shown that CNV is an additional and important source of genetic variability within ADME genes that has significant implications for drug response and precision therapy. In this study we identified a large number of CNVs within ADME genes, many of which had predicted functional impacts. Additionally, many of the CNVs were classified as novel, and several were validated across multiple tools. The CNV landscape within ADME genes was found to be diverse across Africa,

with many variants unique to certain populations, suggesting that no one African population can be used as a proxy for other African populations. However, several of the CNVs identified were common to all SSA regions and had potential functional impacts that warrant further investigation. While a large proportion of the CNVs were common and thus may be useful for the application of precision public health, rare variants with potential deleterious impacts are also important to consider for individual-level precision medicine. Due to the disproportionally high prevalence of disease in SSA, a large proportion of the population are reliant on medication for treatment. By focussing on the drugs most commonly used for the treatment of highly prevalent diseases in the region, we were able to identify several CNVs that are potentially important in an African context. Further investigation of these CNVs is required in order to more thoroughly assess their impact, after which they may be useful to include in future pharmacogenomic tests in order to increase the efficacy of drugs and reduce ADRs, thus providing patients with the best possible treatment outcomes. Overall, this work reinforces the need to consider the unique CNV landscape in African populations when informing safety, efficacy and accurate dosages in Africa. A paper based on this work is currently being drafted to submit for publication and the results will be made available in a suitable database.

2.5 Reference list

Abdullahi, S.T., Soyinka, J.O., Olagunju, A., *et al.* (2020) CYP2B6*6 Genotype Specific Differences in Artemether-Lumefantrine Disposition in Healthy Volunteers. *The Journal of Clinical Pharmacology*. 60(3), pp. 351–360. doi:https://doi.org/10.1002/jcph.1527.

Ahmed, S., Zhou, Z., Zhou, J., *et al.* (2016) Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine. *Genomics, Proteomics and Bioinformatics*. 14(5), pp. 298–313. doi:10.1016/j.gpb.2016.03.008.

Alessandrini, M. and Pepper, M.S. (2014) Priority pharmacogenetics for the African continent: focus on CYP450. *Pharmacogenomics*. 15(3), pp. 385–400. doi:10.2217/pgs.13.252.

Ali, S., Najmi, M.H., Tarning, J., *et al.* (2010) Pharmacokinetics of artemether and dihydroartemisinin in healthy Pakistani male volunteers treated with artemether-lumefantrine. *Malaria Journal*. 9, pp. 275. doi:10.1186/1475-2875-9-275.

Allocati, N., Masulli, M., Di Ilio, C., *et al.* (2018) Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases. *Oncogenesis*. 7(1), pp. 8. doi:10.1038/s41389-017-0025-3.

Avert (2020) *HIV and AIDS in East and Southern Africa Regional Overview*. Available at: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/overview (Accessed: 3 November 2019).

Bains, R.K. (2013) African variation at Cytochrome P450 genes. *Evolution, Medicine, and Public Health*. 2013(1), pp. 118–134. doi:10.1093/emph/eot010.

Bibi, Z. (2008) Role of cytochrome P450 in drug interactions. *Nutrition & Metabolism*. 5(1), pp. 27. doi:10.1186/1743-7075-5-27.

Botton, M.R., Lu, X., Zhao, G., *et al.* (2019) Structural variation at the CYP2C locus: Characterization of deletion and duplication alleles. *Human Mutation*. 40(11), pp. 37–51. doi:10.1002/humu.23855.

Brito, T.C., Possuelo, L.G., Valim, A.R.M., *et al.* (2014) Polymorphisms in CYP2E1, GSTM1 and GSTT1 and anti-tuberculosis drug-induced hepatotoxicity. *Anais da Academia Brasileira de Ciências*. 86, pp. 855–865. doi: 10.1590/0001-3765201420130350.

Buysse, K., Delle Chiaie, B., Van Coster, R., *et al.* (2009) Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *European Journal of Medical Genetics*. 52(6), pp. 398–403. doi:10.1016/j.ejmg.2009.09.002.

Cabianca, D.S. and Gabellini, D. (2010) The cell biology of disease: FSHD: copy number variations on the theme of muscular dystrophy. *The Journal of Cell Biology*. 191(6), pp. 1049–1060. doi:10.1083/jcb.201007028.

Campbell, M.C. and Tishkoff, S.A. (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Hhuman Genetics*. 9, pp. 403–433. doi:10.1146/annurev.genom.9.081307.164258.

Carter, N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*. 39(7), pp. 16–21. doi:10.1038/ng2028.

Chen, X., Schulz-Trieglaff, O., Shaw, R., *et al.* (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 32(8), pp. 1220–1222. doi:10.1093/bioinformatics/btv710.

Cho, D.Y., Shen, J.H.Q., Lemler, S.M., *et al.* (2016) Rifampin enhances cytochrome P450 (CYP) 2B6-mediated efavirenz 8-hydroxylation in healthy volunteers. *Drug Metabolism and Pharmacokinetics*. 31(2), pp. 107–116.

doi:10.1016/j.dmpk.2015.07.002.
Choudhury, A., Aron, S., Sengupta, D., *et al.* (2018) African genetic diversity provides novel insights into evolutionary history and local adaptations. *Human Molecular Genetics*. 27, pp. 209–218. doi:10.1093/hmg/ddy161.
Ciccacci, C., Latini, A., Politi, C., *et al.* (2017) Impact of glutathione transferases genes polymorphisms in nevirapine adverse reactions: a possible role for GSTM1 in SJS/TEN susceptibility. *European Journal of Clinical Pharmacology*. 73(10), pp. 1253–1259. doi:10.1007/s00228-017-2295-2.
Coutelier, M., Holtgrewe, M., Jäger, M., *et al.* (2021) Combining callers improves the detection of copy number variants from whole-genome sequencing. *European Journal of Human Genetics*. 30, pp. 178–186. doi:10.1038/s41431-021-00983-x.
Crews, K.R., Hicks, J.K., Pui, C.H., *et al.* (2012) Pharmacogenomics and individualized medicine: translating science into practice. *Clinical Pharmacology and Therapeutics*. 92(4), pp. 467–475. doi:10.1038/clpt.2012.120.
Cronin-Fenton, D.P., Damkier, P. and Lash, T.L. (2014) Metabolism and transport of tamoxifen in relation to its effectiveness: new perspectives on an ongoing controversy. *Future Oncology*. 10(1), pp. 107–122. doi:10.2217/fon.13.168.
Dandara, C., Swart, M., Mpeta, B., *et al.* (2014) Cytochrome P450 pharmacogenetics in African populations: implications for public health. *Expert Opinion on Drug Metabolism & Toxicology*. 10(6), pp. 769–785. doi:10.1517/17425255.2014.894020.
Dean, L. and Kane, M. (2016) 'Capecitabine Therapy and DPYD Genotype' in Pratt, V.M., Scott, S.A., and Pirmohamed, M., *et al*. (eds.) *Medical Genetics Summaries* [Internet]. Bethesda: National Center for Biotechnology Information.
DecodeGenetics (2019) *svimmer*. Available at: https://github.com/DecodeGenetics/svimmer (Accessed: 22 June 2021).
Dyer, O. (2017) Pfizer and Cipla plan to make cancer drugs affordable in Africa. *BMJ*. 359, pp. 1, doi:10.1136/bmj.j4848.
Eggertsson, H.P., Kristmundsdottir, S., Beyter, D., *et al.* (2019) GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*. 10(1), pp. 5402. doi:10.1038/s41467-019-13341-9.
Elliot, E., Mahungu, T. and Owen, A. (2017) 'Current Progress in the Pharmacogenetics of Infectious Disease Therapy' in Tibayrenc, M. (ed.) *Genetics and Evolution of Infectious Disease*. London: Elsevier. pp. 435–457. doi:https://doi.org/10.1016/B978-0-12-799942-5.00019-6.
Gabrielaite, M., Torp, M.H., Rasmussen, M.S., *et al.* (2021) A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers.* 13(24), pp. 1–21. doi:2021.04.30.442110. doi:10.1101/2021.04.30.442110.
Gaedigk, A. and Leeder, J.S. (2012) CYP2D6, SULT1A1 and UGT2B17 copy number variation: quantitative detection by multiplex PCR. *Pharmacogenomics.* 13, pp. 91–111. doi: 10.2217/pgs.11.135.
Garte, S., Gaspari, L., Alexandrie, A.K., *et al.* (2001) Metabolic gene polymorphism frequencies in control populations. *Cancer Epidemiology, Biomarkers & Prevention*. 10(12), pp. 1239–1248.
Gonzaga-Jauregui, C., Lupski, J. and Gibbs, R. (2012) Human Genome Sequencing in Health and Disease. *Annual Review of Medicine*. 63, pp. 35–61. doi:10.1146/annurev-med-051010-162644.
Handsaker, R.E., Van Doren, V., Berman, J.R., *et al.* (2015) Large multiallelic copy number variations in humans. *Nature Genetics*. 47(3), pp. 296–303. doi:10.1038/ng.3200.

He, X. and Feng, S. (2015) Role of Metabolic Enzymes P450 (CYP) on Activating Procarcinogen and their Polymorphisms on the Risk of Cancers. *Current Drug Metabolism*. 16(10), pp. 850–863. doi:10.2174/1389200216101512101645O1.
He, Y., Hoskins, J.M. and McLeod, H.L. (2011) Copy number variants in pharmacogenetic genes. *Trends in Molecular Medicine*. 17(5), pp. 244–251. doi:10.1016/j.molmed.2011.01.007.
Hovelson, D.H., Xue, Z., Zawistowski, M., *et al.* (2017) Characterization of ADME gene variation in 21 populations by exome sequencing. *Pharmacogenetics and Genomics*. 27(3), pp. 89–100. doi:10.1097/FPC.0000000000000260.
Howe, K.L., Achuthan, P., Allen, James, *et al.* (2021) Ensembl 2021. *Nucleic Acids Research*. 49, pp. 884–891. doi:10.1093/nar/gkaa942.
Ilett, K.F., Ethell, B.T., Maggs, J.L., *et al.* (2002) Glucuronidation of dihydroartemisinin in vivo and by human liver microsomes and expressed UDP-glucuronosyltransferases. *Drug Metabolism and Disposition*. 30(9), pp. 1005–1012. doi:10.1124/dmd.30.9.1005.
Ingelman-Sundberg, M., Mkrtchian, S., Zhou, Y., *et al.* (2018) Integrating rare genetic variants into pharmacogenetic drug response predictions. *Human Genomics*. 12(1), p. 26. doi:10.1186/s40246-018-0157-3.
Jakubosky, D., Smith, E.N., D'Antonio, M., *et al.* (2020) Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. *Nature Communications*. 11(1), pp. 2928. doi:10.1038/s41467-020-16481-5.
Jarvis, J.P., Peter, A.P. and Shaman, J.A. (2019) Consequences of CYP2D6 Copy-Number Variation for Pharmacogenomics in Psychiatry. *Frontiers in Psychiatry*. 10, pp. 432. doi:10.3389/fpsyt.2019.00432.
Johansson, I. and Ingelman-Sundberg, M. (2009) CNVs of human genes and their implication in pharmacogenetics. *Cytogenetic and Genome Research*. 123(1–4), pp. 195–204. doi:10.1159/000184709.
Kenmotsu, H. and Tanigawara, Y. (2015) Pharmacokinetics, dynamics and toxicity of docetaxel: Why the Japanese dose differs from the Western dose. *Cancer Science*. 106(5), pp. 497–504. doi:10.1111/cas.12647.
Kis, O., Sankaran-Walters, S., Hoque, M.T., *et al.* (2016) HIV-1 Alters Intestinal Expression of Drug Transporters and Metabolic Enzymes: Implications for Antiretroviral Drug Disposition. *Antimicrobial Agents and Chemotherapy*. 60(5), pp. 2771–2781. doi:10.1128/AAC.02278-15.
Korhonova, M., Doricakova, A. and Dvorak, Z. (2015) Optical Isomers of Atorvastatin, Rosuvastatin and Fluvastatin Enantiospecifically Activate Pregnane X Receptor PXR and Induce CYP2A6, CYP2B6 and CYP3A4 in Human Hepatocytes. *PLoS One*. 10(9), pp. 1–18. doi:10.1371/journal.pone.0137720.
Kosugi, S., Momozawa, Y., Liu, X., *et al.* (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*. 20(1), pp. 117. doi:10.1186/s13059-019-1720-5.
Langmia, I.M., Just, K.S., Yamoune, S., *et al.* (2021) CYP2B6 Functional Variability in Drug Metabolism and Exposure Across Populations – Implication for Drug Safety, Dosing, and Individualized Therapy. *Frontiers in Genetics.* 12(12), pp. 1–21. doi: 10.3389/fgene.2021.692234.
Layer, R.M., Chiang, C., Quinlan, A.R., *et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*. 15(6), pp. 84. doi:10.1186/gb-2014-15-6-r84.
Lim, L. and Burge, C. (2001) A computational analysis of sequence features involved in recognition of short introns. *PNAS.* 98(20), pp. 11193–11198.
MacDonald, J.R., Ziman, R., Yuen, R.K.C., *et al.* (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic*

Acids Research. 42, pp. 986–992. doi:10.1093/nar/gkt958.
Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., *et al.* (2019) Structural variant calling: the long and the short of it. *Genome Biology*. 20(1), pp. 246. doi:10.1186/s13059-019-1828-7.
Marcath, L.A., Kidwell, K.M., Robinson, A.C., *et al.* (2019) Patients carrying CYP2C8*3 have shorter systemic paclitaxel exposure. *Pharmacogenomics*. 20(2), pp. 95–104. doi:10.2217/pgs-2018-0162.
Martin, A.R., Teferra, S., Möller, M., *et al.* (2018) The critical needs and challenges for genetic architecture studies in Africa. *Current Oinion in Genetics & Development*. 53, pp. 113–120. doi:10.1016/j.gde.2018.08.005.
Martis, S., Mei, H., Vijzelaar, R., *et al.* (2013) Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. *The Pharmacogenomics Journal*. 13(6), pp. 558–566. doi:10.1038/tpj.2012.48.
Mayosi, B.M., Flisher, A.J., Lalloo, U.G., *et al.* (2009) The burden of non-communicable diseases in South Africa. *Lancet*. 374(9693), pp. 934–947. doi:10.1016/S0140-6736(09)61087-4.
Mittal, B., Tulsyan, S., Kumar, S., *et al.* (2015) 'Cytochrome P450 in Cancer Susceptibility and Treatment' in Makowski, G. (ed.) *Advances in Clinical Chemistry*. Amsterdam: Elsevier. pp. 77–139. doi:https://doi.org/10.1016/bs.acc.2015.06.003.
Mulder, N., Lombard, Z., Owolabi, M.O., *et al.* (2020) The Genetic and Environmental Basis for Diseases in Understudied Populations. *Frontiers in Genetics*. 11(559956), pp. 1–2. doi: 10.3389/fgene.2020.559956.
Navaratnam, V., Mansor, S.M., Sit, N.W., *et al.* (2000) Pharmacokinetics of artemisinin-type compounds. *Clinical Pharmacokinetics*. 39(4), pp. 255–270. doi:10.2165/00003088-200039040-00002.
Nowakowska, B. (2017) Clinical interpretation of copy number variants in the human genome. *Journal of Applied Genetics*. 58(4), pp. 449–457. doi:10.1007/s13353-017-0407-4.
Nyirenda, M.J. (2016) Non-communicable diseases in sub-Saharan Africa: Understanding the drivers of the epidemic to inform intervention strategies. *International Health*. 8(3), pp. 157–158. doi:10.1093/inthealth/ihw021.
Olsson, A.G., McTaggart, F. and Raza, A. (2002) Rosuvastatin: a highly effective new HMG-CoA reductase inhibitor. *Cardiovascular Drug Reviews*. 20(4), pp. 303–328. doi:10.1111/j.1527-3466.2002.tb00099.x.
Pabinger, S., Dander, A., Fischer, M., *et al.* (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*. 15(2), pp. 256–278. doi:10.1093/bib/bbs086.
PharmaADME.org (2021) *PharmaADME.org*. Available at: http://pharmaadme.org. (Accessed: 15 June 2019)
Piacentini, S., Polimanti, R., Porreca, F., *et al.* (2011) GSTT1 and GSTM1 gene polymorphisms in European and African populations. *Molecular Biology Reports*. 38(2), pp. 1225–1230. doi:10.1007/s11033-010-0221-0.
Pirooznia, M., Goes, F. and Zandi, P.P. (2015) Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics*. 6(138), pp. 1–9. doi:10.3389/fgene.2015.00138.
Png, G., Suveges, D., Park, Y., *et al.* (2019) Population-wide copy number variation calling using variant call format files from 6,898 individuals. *Genetic Epidemiology*. 44(1), pp. 79–89. doi:10.1002/gepi.22260.
Poplin, R., Ruano-Rubio, V., DePristo, M.A., *et al.* (2018) Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. doi:10.1101/201178.

Ramírez, B., Niño-Orrego, M.J., Cárdenas, D., *et al.* (2019) Copy number variation profiling in pharmacogenetics CYP-450 and GST genes in Colombian population. *BMC Medical Genomics*. 12(1), pp. 110. doi:10.1186/s12920-019-0556-x.

Rao, Y., Hoffmann, E., Zia, M., *et al.* (2000) Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. *Molecular Pharmacology*. 58(4), pp. 747–755. doi:10.1124/mol.58.4.747.

Rausch, T., Zichner, T., Schlattl, A., *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 28(18), pp. 333–339. doi:10.1093/bioinformatics/bts378.

Rigau, M., Juan, D., Valencia, A., *et al.* (2019) Intronic CNVs and gene expression variation in human populations. *PLoS Genetics*. 15(1), pp. 1–23. doi:10.1371/journal.pgen.1007902.

Saif, M.W., Ezzeldin, H., Vance, K., *et al.* (2007) DPYD*2A mutation: the most common mutation associated with DPD deficiency. *Cancer Chemotherapy and Pharmacology*. 60(4), pp. 503–507. doi:10.1007/s00280-006-0392-5.

Saif, M.W., Ezzeldin, H., Vance, K., *et al.* (2019) Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Nucleic Acids Research*. 123(1), pp. 333–342. doi:10.1038/s41436-018-0033-5.

Saini, R., Saini, S. and Sugandha, R. (2010) Pharmacogenetics: The future medicine. *Journal of Advanced Pharmaceutical Technology & Research*. 1(4), pp. 423–424. doi:10.4103/0110-5558.76443.

Santos, M., Niemi, M., Hiratsuka, M., *et al.* (2018) Novel copy-number variations in pharmacogenes contribute to interindividual differences in drug pharmacokinetics. *Genetics in Medicine*. 20(6), pp. 622–629. doi:10.1038/gim.2017.156.

Schipani, A., Wyen, C., Mahungu, T., *et al.* (2011) Integration of population pharmacokinetics and pharmacogenetics: an aid to optimal nevirapine dose selection in HIV-infected individuals. *Journal of Antimicrobial Chemotherapy*. 66(6), pp. 1332–1339. doi:10.1093/jac/dkr087.

Scripture, C.D. and Pieper, J.A. (2001) Clinical pharmacokinetics of fluvastatin. *Clinical Pharmacokinetics*. 40(4), pp. 263–281. doi:10.2165/00003088-200140040-00003.

Sim, S.C., Kacevska, M. and Ingelman-Sundberg, M. (2012) Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. *The Pharmacogenomics Journal*. 13, pp. 1–11. doi: 10.1038/tpj.2012.45.

de Smith, A.J., Walters, R.G., Froguel, P., *et al.* (2009) Human genes involved in copy number variation: Mechanisms of origin, functional effects and implications for disease. *Cytogenetic and Genome Research*. 123(1–4), pp. 17–26. doi:10.1159/000184688.

Spratlin, J. and Sawyer, M.B. (2007) Pharmacogenetics of paclitaxel metabolism. *Critical Reviews in Oncology/Hematology*. 61(3), pp. 222–229. doi:https://doi.org/10.1016/j.critrevonc.2006.09.006.

Stranger, B.E., Forrest, M.S., Dunning, M., *et al.* (2007) Relative Impact of Nucleotide and Copy Number Variation on Gene Expression. *Science*. 315(5813), pp. 848–853. doi:10.1126/science.1137223.

Takahashi, H. and Echizen, H. (2003) Pharmacogenetics of CYP2C9 and interindividual variability in anticoagulant response to warfarin. *The Pharmacogenomics Journal*. 3(4), pp. 202–214. doi:10.1038/sj.tpj.6500182.

Tremmel, R., Klein, K., Battke, F., *et al.* (2020) Copy number variation profiling in pharmacogenes using panel-based exome resequencing and correlation to human liver expression. *Human Genetics*. 139(2), pp. 137–149. doi:10.1007/s00439-019-02093-7.

Tremmel, R., Klein, K., Winter, S., *et al.* (2016) Gene copy number variation analysis reveals dosage-insensitive expression of CYP2E1. *The Pharmacogenomics Journal*. 16(6), pp. 551–558. doi:10.1038/tpj.2015.69.

Ugartondo, N., Martínez-Gil, N., Esteve, M., *et al.* (2021) Functional Analyses of Four CYP1A1 Missense Mutations Present in Patients with Atypical Femoral Fractures. *International Journal of Molecular Sciences*. 22(14), pp. 1–11. doi:10.3390/ijms22147395.

Vijzelaar, R., Botton, M.R., Stolk, L., *et al.* (2018) Multi-ethnic SULT1A1 copy number profiling with multiplex ligation-dependent probe amplification. *Pharmacogenomics*. 19(9), pp. 761–770. doi:10.2217/pgs-2018-0047.

Wang, P., Pradhan, K., Zhong, X.B., *et al.* (2016) Isoniazid metabolism and hepatotoxicity. *Acta pharmaceutica Sinica B*. 6(5), pp. 384–392. doi:10.1016/j.apsb.2016.07.014.

World Health Organization (2020a) *Malaria*. Available at: https://www.who.int/news-room/fact-sheets/detail/malaria (Accessed: 30 April 2020).

World Health Organization (2020b) *Tuberculosis*. Available at: https://www.who.int/health-topics/tuberculosis#tab=tab_1 (Accessed: 30 April 2020).

Yan, M.M., Wu, S.S., Ying, Y.Q., *et al.* (2018) Safety assessment of concurrent statin treatment and evaluation of drug interactions in China. *SAGE Open Medicine*. 6, pp. 1–9. doi:10.1177/2050312118798278.

Zang, M., Zhu, F., Zhao, L., *et al.* (2014) The effect of UGTs polymorphism on the auto-induction phase II metabolism-mediated pharmacokinetics of dihydroartemisinin in healthy Chinese subjects after oral administration of a fixed combination of dihydroartemisinin-piperaquine. *Malaria Journal*. 13(1), pp. 478. doi:10.1186/1475-2875-13-478.

Zhang, L., Bai, W., Yuan, N., *et al.* (2019) Comprehensively benchmarking applications for detecting copy number variation. *PLoS Computational Biology*. 15(5), pp. 1–12. doi:10.1371/journal.pcbi.1007069.

Zhao, M., Wang, Qingguo, Wang, Quan, *et al.* (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 14(11), pp. 1–16. doi:10.1186/1471-2105-14-S11-S1.

Zhou, X.Y., Hu, X.X., Wang, C.C., *et al.* (2019) Enzymatic Activities of CYP3A4 Allelic Variants on Quinine 3-Hydroxylation In Vitro. *Frontiers in pharmacology*. 10, pp. 591. doi:10.3389/fphar.2019.00591.

CHAPTER 3: VALIDATION OF SHORT-READ CNVS USING A LONG-READ SEQUENCING APPROACH

3. CHAPTER 3: VALIDATION OF SHORT-READ CNVS USING A LONG-READ SEQUENCING APPROACH

3.1 Introduction

In recent years, various studies have uncovered the important role that structural variation (SV) plays in the human genome. However, our understanding of SV, including copy number variants (CNVs), has been limited by the available technology (Sedlazeck *et al.*, 2018). The majority of studies on CNV have made use of short-read sequencing technologies, which produce reads that are on average 200 bp in length (Sedlazeck *et al.*, 2018; Hon *et al.*, 2020). When these short reads are aligned to the reference genome, incomplete assemblies with sequence gaps are produced (*Figure 3.1*). This has been shown to be problematic for SV calling, and the results obtained from short-read sequencing data suffer from low sensitivity (30-70%) and high false discovery rates, as high as 85% (Sedlazeck *et al.*, 2018).

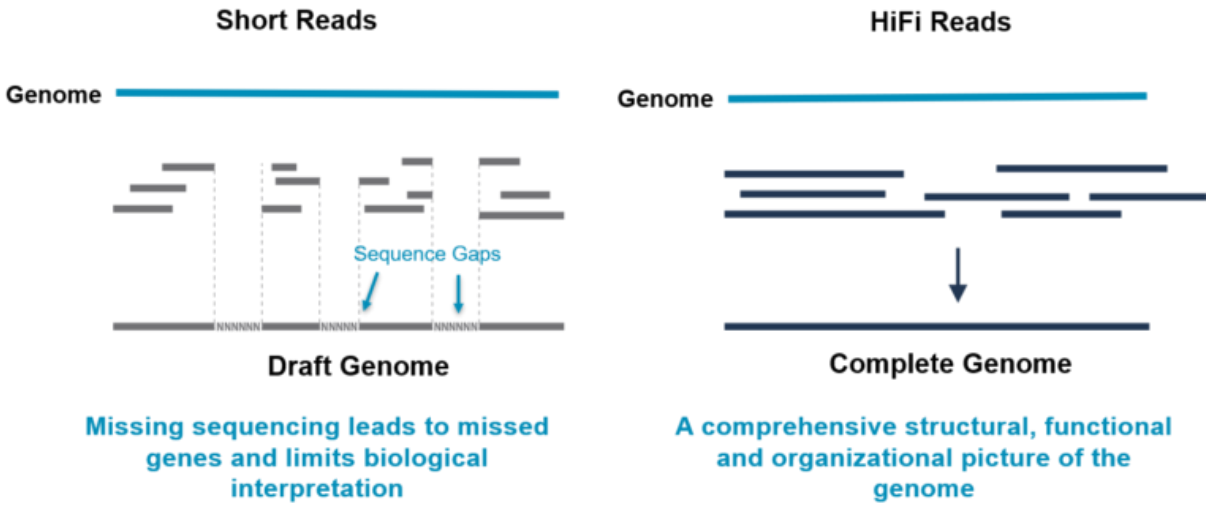


Figure 3.1: Comparison of short-read and Pacific Biosciences (PacBio) HiFi sequencing approaches (PacBio, 2020).

Recently, PacBio has developed an approach that makes use of multiple pass circular consensus sequencing of individual molecules, producing highly-accurate long reads (High Fidelity (HiFi) reads) (Wenger *et al.*, 2019).

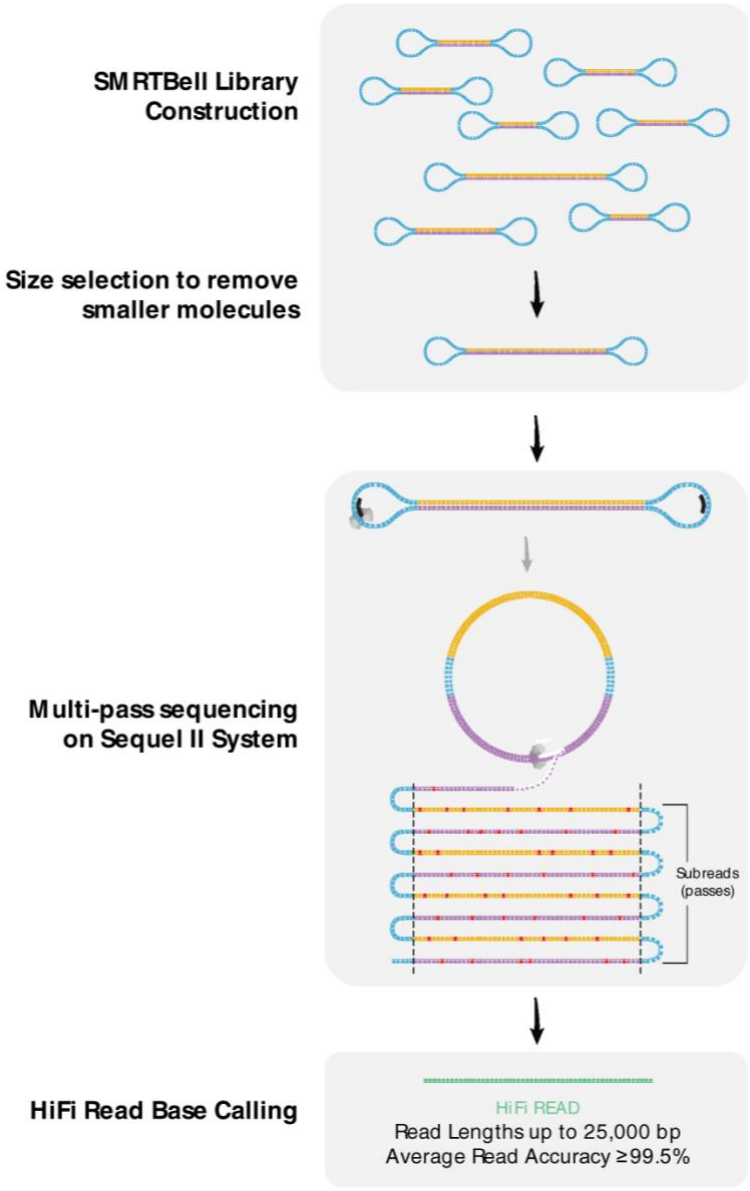


Figure 3.2: Flowchart of the PacBio HiFi sequencing approach (Wenger *et al*. 2019, licensed under Creative Commons CC BY 4.0 https://creativecommons.org/licenses/by/4.0/. No changes have been made to the figure content).

The process of HiFi sequencing is as follows (*Figure 3.2*): Following DNA isolation, a SMRTBell library is generated by ligating adaptors to each end of double-stranded DNA creating a circular template called a linear SMRTBell template. Primer and polymerase are added to the library, and after the sequencing primer has annealed, the polymerase can bind to the template. The library is then loaded onto the instrument for sequencing. The main component of this approach is the single-molecule real-time (SMRT) cell, which contains millions of cells called zero-mode wave guides (ZMWs). A single molecule of DNA is immobilized in each ZMW. The sequencing polymerase

has strand-displacing capabilities, and will open the linear SMRTBell template, to form a circular structure. As the polymerase incorporates labelled nucleotides, light is emitted, and with this approach, nucleotide incorporation is measured in real-time. The polymerase incorporates labelled nucleotides in a rolling-circle manner, which enables the sequencing of both the forward and reverse strands and will continue sequencing around the SMRTBell template until the sequencing run ends. With every pass around the template, a consensus can be generated, and as any sequencing errors are randomly distributed, they become diluted by increasing passes around the circular template. While previous long-read technologies have been error-prone, this approach has significantly increased the accuracy of the reads generated.

This sequencing approach produces reads that are 10–25 kb in length on average, with accuracies greater than 99.5% (Hon *et al.*, 2020). When these long reads are aligned, they create overlaps that allow for the generation of complete genome assemblies (*Figure 3.1*). This, in combination with the fact that long reads have the ability to span entire SVs, make this approach considerably more accurate for SV calling. Wenger and colleagues showed that HiFi sequencing exceeded the ability of short-read sequencing approaches to detect SVs, and obtained 96.13% precision and 95.99% recall for deletions and duplications using this sequencing approach, in combination with pbsv for SV calling. (Wenger *et al.*, 2019).

It has been shown that using a long-read sequencing approach increases the number of SVs detected. For example, the 1000 Genomes Project (1KGP) reported 69 000 SVs based on the analysis of 2 504 short-read whole genome sequences (Ebert *et al.*, 2021). However, a recent study that made use of PacBio long-read sequencing, in combination with single-cell DNA template strand sequencing (Strand-seq), on only 32 samples identified 107 126 SVs, which more than tripled the rate of discovery when compared with short-read SV analysis on the same samples (Ebert *et al.*, 2021). Furthermore, recent SV studies making use of thousands of short-read sequences have reported the number of SVs per sample to be between 5 000 and 10 000 (Abel *et al.*, 2020; Collins *et al.*, 2020). On the other hand, long-read approaches have identified between 23 000 and 28 000 SVs per sample (Ebert *et al.*, 2021). The predominant increase in the number of variants identified using long-read approaches can be attributed to small SVs (< 250 bp in length) (Ebert *et al.*, 2021).

The aims of the following work are to use PacBio long-read sequencing to:

(i) validate the CNVs called from short-read sequence data;

(ii) gain a deeper understanding of the CNV landscape; and

(iii) assess whether short-read sequencing is adequate for accurate CNV calls, or whether long-read sequencing is required.

3.2 Materials and Methods

3.2.1 Study participants

Eight samples, obtained from South African individuals enrolled in the AWI-Gen study (Ramsay *et al.*, 2016), were selected for whole-genome long-read sequencing. The samples were chosen based on self-reported ethno-linguistic categories in order to sample individuals from different South Eastern Bantu-speaking groups, as well as availability of DNA. A minimum of 15 µg of DNA was required for sequencing, therefore samples were chosen based on the availability of DNA samples with adequate amounts of DNA available. The breakdown of self-reported ethno-linguistic categories were as follows: Tsonga (AB0330 and AB2543); Tswana (CVN0L and DFX0S); Venda (AJK0F and DEC0V) and Xhosa (CFF0X and CWF0F).

3.2.2 HiFi sequencing

HiFi sequencing was performed by Inqaba Biotech using the following protocol. **(1) gDNA quality control:** the gDNA quality was assessed on the TapeStation using Genomic DNA ScreenTape and the gDNA was quantified on the Qubit using the Qubit HS dsDNA Assay kit. **(2) gDNA Shearing and Library preparation:** gDNA was sheared to ~10–15 kb fragments using g-TUBEs from Covaris. The samples were purified with AMPure PB beads (for size selection) and were quantified and assessed on the TapeStation. The HiFi sequencing library was prepared using the 'PacBio HiFi SMRTBell Libraries using SMRTBell Express Template Prep Kit 2.0' as per the standard protocol: https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf. (3) **Library quality control:** The quality of the HiFi SMRTBell library was assessed using the TapeStation and quantified using the Qubit. **(4) Sequencing: '**Sample Setup' and 'Run Design' were performed using SMRTLink software (v10.1) and subsequent sequencing was performed on the PacBio Sequel IIe system using SMRT Cells 8M. Each sample was sequenced on two SMRT Cells, with the exception of CFF0X which was sequenced on an additional SMRT Cell, as the original quality was unsatisfactory.

3.2.3 Data analysis

Generation of BAM files and SV calling

BAM files were generated by aligning the fastq files containing the HiFi reads to the GRCh38 reference genome. This step was performed using pbmm2 (v1.7.0; https://github.com/PacificBiosciences/pbmm2) with default parameters. While several tools are available for SV calling from long-read sequencing data, a recent study found that pbsv defines SVs most accurately across all datasets and since it is designed for PacBio data, it performs the best on this type data (Dierckxsens *et al*., 2020). Following generation of the BAM files, pbsv discover (v2.6.2; https://github.com/ PacificBiosciences/pbsv) was run with default parameters to produce files containing SV signatures (.svsig files). These files were passed to pbsv call (v2.6.2; https://github.com/ PacificBiosciences/pbsv) which called SVs from the SV signatures using default parameters. The Nextflow script which was used to generate the BAM files and run pbsv, along with the accompanying configuration file, can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%203%20scripts/run_pbmm2_pbsv.nf and https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%203%20scripts/run_pbmm2_pbsv.config)**.** Additionally, pbsv call was also used to call structural variants jointly for all samples of interest, and variant calls for all samples were output in a single VCF file. The bash script used for joint calling is also available in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%203%20scripts/joint_cnv_call.sh).

Quality control

Mosdepth (v0.3.2; https://github.com/brentp/mosdepth/releases/tag/v0.3.2) was used to calculate the coverage of the BAM files over 5 kb windows. The coverage information from the mosdepth files was passed to an R script which was used to generate coverage plots for each BAM file.

As pbsv calls all SV types and we are only interested in deletions and duplications, we filtered the SVTYPE INFO field to include only deletions and duplications. Only variants that obtained a PASS for the FILTER value and those that were greater than 50 bp in size were included. All of the above filtering was performed using BCFtools' view function (v1.10.2). Additionally, variants that were supported by fewer than five reads

(obtained from the DP field which specifies the read depth for each variant for each sample) were excluded from any downstream analysis. This filtering was performed using GATK's (v4.1.8.1) VariantFiltration function, whereby for sites where the read depth was less than five, a 'DP-FAIL' tag was added and the genotype was set to no call (./.). Following this, GATK's (v4.1.8.1) SelectVariants function was used to exclude non-variant sites (i.e. those sites that have no genotype (./.) or only 0/0 genotypes across all samples).

CNV annotation

Basic Linux commands as well as VCF manipulating tools such as BCFtools (v1.10.2) and VCFtools (v0.1.16) were used for data analysis. Ensembl's Variant Effect Predictor (VEP) (v104) (Howe *et al.*, 2021) was used to determine which genes were overlapped by CNVs, as well as to assess the functional consequences of the CNVs. Variants were compared to those in the Database of Genomic Variation (DGV) using BEDTools intersect (v2.27.1) in order to determine whether they had been previously identified by other studies or whether they were novel. The CNVs identified using long-read sequencing were compared to those identified using a short-read sequencing approach in Chapter 2 using BEDTools intersect (v2.27.1) with 50% reciprocal overlap. A file containing the number of variants per sample was generated using Genome STRiP's (svtoolkit_2.00.1918) (Handsaker *et al.*, 2015) SVAnnotator pipeline (https://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_main_SVAnnotator.html). All graphs were plotted using Microsoft Excel (v16.54) or R Studio (UpSetR package) (v1.2.1335).

3.3 Results

3.3.1 Sequencing quality

Various metrics were used to assess the quality of the sequencing reads (*Table 3.1*). Each sample was run on two SMRT cells, with the exception of CFF0X. The overall quality of CFF0X was unsatisfactory, therefore, this sample was subsequently rerun on an additional SMRT cell. The number of HiFi reads generated per cell ranged from 621 691 to 3 288 105, with an average of 2 038 346 HiFi reads per cell. The mean HiFi read length was 10 429 bp across all 17 SMRT cells, and the median read quality ranged from Q31 to Q39, with an average of Q35. Phred quality is a measure of the quality of base calling, whereby the higher the Phred score, the lower the probability of a base calling error and the higher the base calling accuracy. For example, a Phred score of Q35 indicates that the probability of an incorrect base call is 0.00031623, and therefore the base calling accuracy is 99.96837722%.

Once the reads had been aligned to the reference genome to generate BAM files, the coverage of each sample was assessed. The coverage plots for each sample are available in Appendix B (*Figure B1-10*). The coverage across the genome ranged from 10.6x to 18.8x across the eight samples. The original coverage of CFF0X was only 4.7x (*Figure B4*), with just under half (49.3%) of the genome having low (less than 5x coverage) to no coverage. Due to the quality concerns of this sample, it was rerun on an additional SMRT cell, and the second BAM file generated had an average coverage across the genome of 5.9x (*Figure B5*). When the data from the three SMRT cells was combined, an improved coverage of 10.6x was obtained for this sample (*Figure B6*).

	AB0330		AB2543		AJK0F		CFF0X		
	Cell 1	**Cell 2**	**Cell 1**	**Cell 2**	**Cell 1**	**Cell 2**	**Cell 1**	**Cell 2**	**Cell 3**
HiFi Reads	3 110 271	2 660 324	2 431 757	2 630 074	2 210 802	2 266 344	621 691	704 090	1 849 312
HiFi Yield (bp)	32 333 707 215	28 585 735 303	24 453 710 276	25 966 477 497	23 288 102 306	23 795 805 300	7 184 915 960	7 996 128 188	18 216 002 136
HiFi Read Length (mean, bp)	10 395	10 745	10 055	9 872	10 533	10 499	11 557	11 356	10 390
HiFi Read Quality (median)	Q36	Q32	Q37	Q37	Q31	Q32	Q35	Q34	Q33
HiFi Number of Passes (mean)	14	14	14	15	12	12	13	13	13
<Q20 Reads	440 231	515 419	341 919	365 115	571 361	545 745	77 097	92 556	376 887
<Q20 Yield (bp)	4 919 700 060	5 999 243 139	3 729 737 765	3 891 851 994	6 976 358 729	6 596 423 761	967 082 484	1 150 451 981	4 092 232 270
<Q20 Read Length (mean, bp)	11 175	11 639	10 908	10 659	12 210	12 087	12 543	12 429	10 857
<Q20 Read Quality (median)	Q16	Q16	Q16	Q16	Q16	Q16	Q16	Q16	Q16
Coverage	18.8x		15.6x		14.5x		10.6x		

Table 3.1: Various quality metrics for the HiFi sequencing of eight samples.

	CVN0L		CWF0F		DEC0V		DSX0S	
	Cell 1	**Cell 2**	**Cell 1**	**Cell 2**	**Cell 1**	**Cell 2**	**Cell 1**	**Cell 2**
HiFi Reads	1 367 604	1 725 426	1 650 603	1 625 965	3 288 105	2 981 630	1 712 796	1 815 085
HiFi Yield (bp)	16 180 625 238	18 274 525 573	16 074 453 517	16 441 456 010	24 349 152 630	27 380 118 404	19 960 970 914	20 668 418 314
HiFi Read Length (mean, bp)	11 831	10 591	9 738	10 111	7 405	9 182	11 654	11 387
HiFi Read Quality (median)	Q36	Q35	Q34	Q37	Q39	Q38	Q35	Q34
HiFi Number of Passes (mean)	14	14	14	14	17	16	13	13
<Q20 Reads	170 265	249 214	247 901	202 646	436 779	366 678	267 398	309 822
<Q20 Yield (bp)	2 212 211 844	3 016 026 995	2 644 033 354	2 199 952 129	3 667 305 751	3 609 284 419	3 376 249 000	3 871 721 021
<Q20 Read Length (mean, bp)	12 992	12 102	10 665	10 856	8 396	9 843	12 626	12 496
<Q20 Read Quality (median)	Q16	Q16	Q16	Q16	Q16	Q16	Q16	Q16
Coverage	10.6x		10x		16x		12.5x	

3.3.2 CNV landscape across the genome

From the eight long-read sequences, 42 336 CNVs were identified across the genome of which 28 518 (67.4%) were deletions and 13 818 (32.6%) were duplications. A file containing these CNVs can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Supplementary%20Files/Long_Read_CNVs.xlsx).

The number of CNVs identified per sample can be seen in *Table 3.2* below. The greatest number of CNVs were identified in AB0330 (n = 17 153), while CFF0X harboured the fewest CNVs (n = 16 042). On average, each sample harboured a total of 16 493 CNVs across the genome. Additionally, a number of variants were identified in only a single sample (i.e. unique variants). This ranged from 1 563 in CFF0X to 2 110 in CVN0L. The average number of unique CNVs per sample was 1 831.

Table 3.2: Number of CNVs and unique CNVs identified in each sample.

Sample	Number of CNVs	Number of unique CNVs
AB0330	17 153	1 870
AB2543	16 700	1 748
AJK0F	16 232	1 741
CFF0X	16 042	1 563
CVN0L	16 096	2 110
CWF0F	16 188	1 752
DEV0V	17 044	1 890
DSX0S	16 487	1 974

We looked at the CNV distribution across the chromosomes in order to assess whether specific chromosomes were more affected by CNVs than others. The number of CNVs identified on each chromosome and the proportion of CNVs scaled by chromosome length can be seen in *Table 3.3* and *Figure 3.3* respectively. Chromosome 2 harbours the highest number of CNVs (n = 3 476), while chromosome 21 harbours the fewest (n = 1 023) (*Table 3.3*). The size of the chromosome does not directly correlate with the number of CNVs identified on the chromosome. For example, chromosome 7 (n = 2 820) harbours a higher proportion of CNVs compared to its larger neighbours, and the smallest chromosomes (16, 17, 18, 19, 20, 21 and 22) have a higher proportion of CNVs compared to nearly all of the larger chromosomes (*Figure 3.3*). Chromosome 14 (n = 1 127) and chromosome 15 (n = 1 061) have the smallest proportion of CNVs compared to the other chromosomes (*Figure 3.3*). Additionally, 5 313 (12.6%) of the CNVs overlapped with regions of the genome containing segmental duplications.

Table 3.3: Number of CNVs (deletions and duplications) on each chromosome.

Chromosome	Number of deletions	Number of duplications	Total number of CNVs
chr1	2 017	913	2 930
chr2	2 384	1 092	3 476
chr3	1 578	678	2 256
chr4	1 899	833	2 732
chr5	1 648	827	2 475
chr6	1 483	743	2 226
chr7	1 866	954	2 820
chr8	1 280	639	1 919
chr9	1 170	509	1 679
chr10	1 433	814	2 247
chr11	1 203	560	1 763
chr12	1 314	638	1 952
chr13	1 204	556	1 760
chr14	802	325	1 127
chr15	735	326	1 061
chr16	1 065	587	1 652
chr17	1 127	665	1 792
chr18	905	444	1 349
chr19	1 044	579	1 623
chr20	936	432	1 368
chr21	670	353	1 023
chr22	755	351	1 106

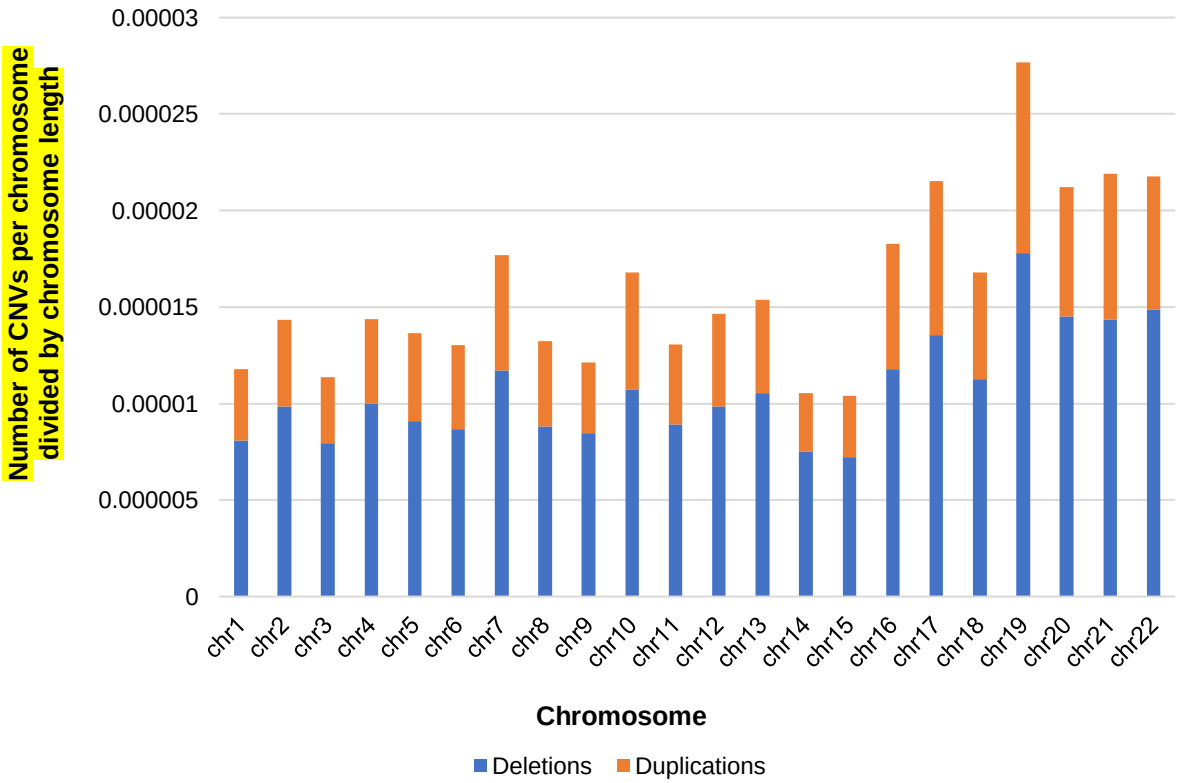


Figure 3.3: Number of deletions (blue) and duplications (orange) identified on each chromosome divided by chromosome length.

A total of 18 157 (42.9%) of CNVs were classified as novel, based on having less than 50% reciprocal overlap with CNVs in the DGV.

The upset plot below (*Figure 3.4*) shows the overlap of CNVs between samples. Just over half (53%) of the CNVs were unique to a sample or were observed in only two of the eight samples. A total of 3 024 (7.1%) CNVs were observed across all eight individuals. Upset plots for the Tsonga samples (AB0330 and AB2543) (*Figure C2*); Tswana samples (CVN0L and DFX0S) (*Figure C1*); Venda samples (AJK0F and DEC0V) (*Figure C3*) and Xhosa samples (CFF0X and CWF0F) (*Figure C4*) can be seen in Appendix C. Analysis of the two Xhosa samples showed that these two samples shared the highest number of CNVs (40.9%). This was followed by the two Tsonga samples which shared 39.1% of CNVs, the two Venda samples which shared 37.44% of CNVs and lastly the two Tswana samples which shared 35.2% of CNVs.

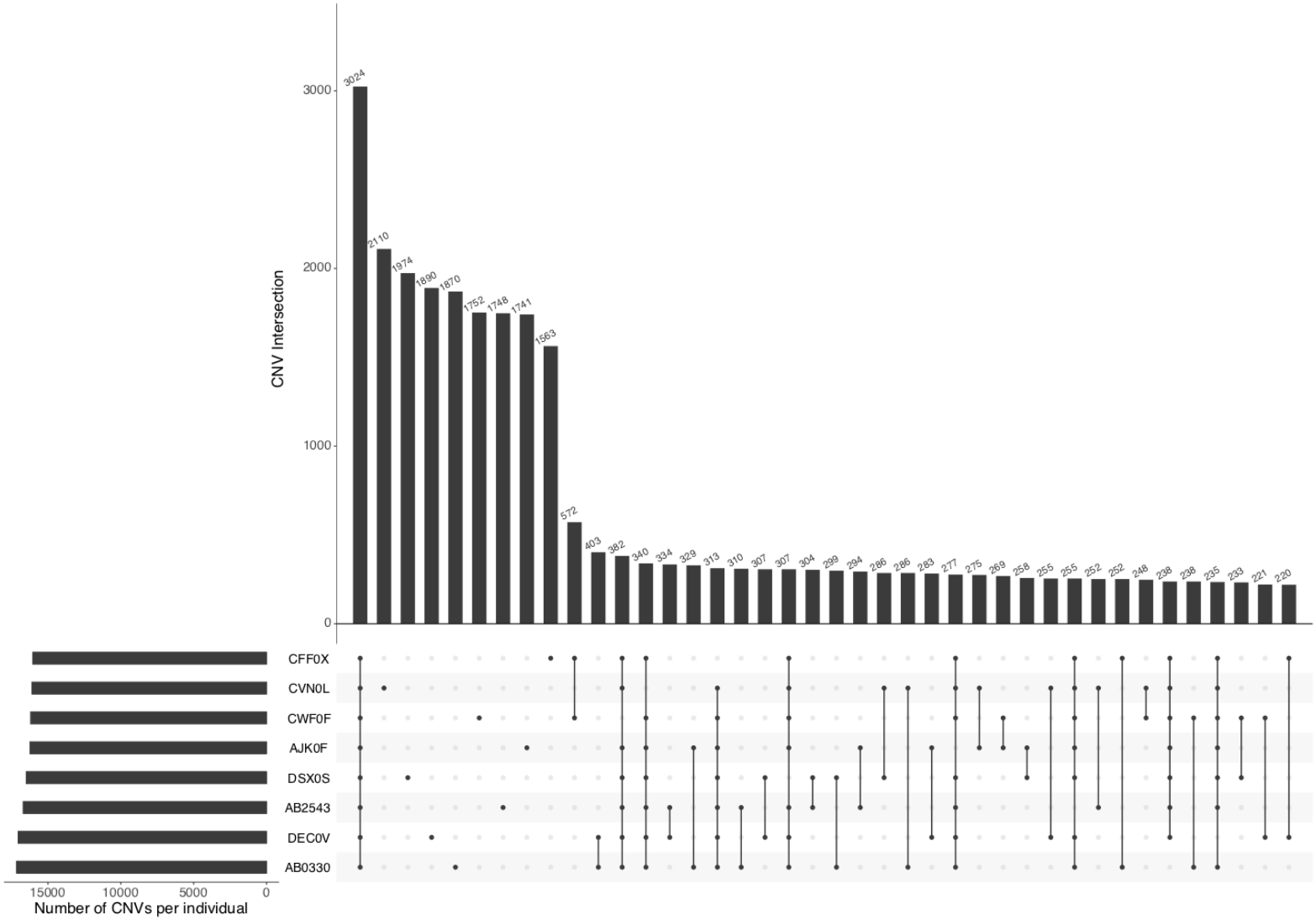


Figure 3.4: Upset plot showing the overlap between the CNVs identified in each sample.

3.3.3 Consequences of CNVs

As the size of a CNV can affect the CNVs functional impact, the size distribution of the CNVs was assessed and has been visualised in *Figure 3.5* below. The mean and median sizes of the CNVs were 517 bp and 134 bp respectively. The minimum size was 50 bp (minimum cut-off size based on CNV definition) and the maximum size was 99 kb. Deletions were found to be larger than duplications, with an average size of 629 bp and 284 bp respectively. The majority of CNVs (91.6%) were small CNVs (<1000 bp). The most common size range of CNVs was 50–100 bp (n = 16 509; 39%), followed by 101–200 bp (n = 10 374; 24.5%). Over 80% of the CNVs were between 50 bp and 500 bp in size. We identified several (n = 3 577; 8.4%) large CNVs which were greater than 1000 bp in size, the majority of which (84%) were deletions. Interestingly, for the 301–400 bp size range we observed an increase in the number of deletions which accounted for 77% of the CNVs in that size range.

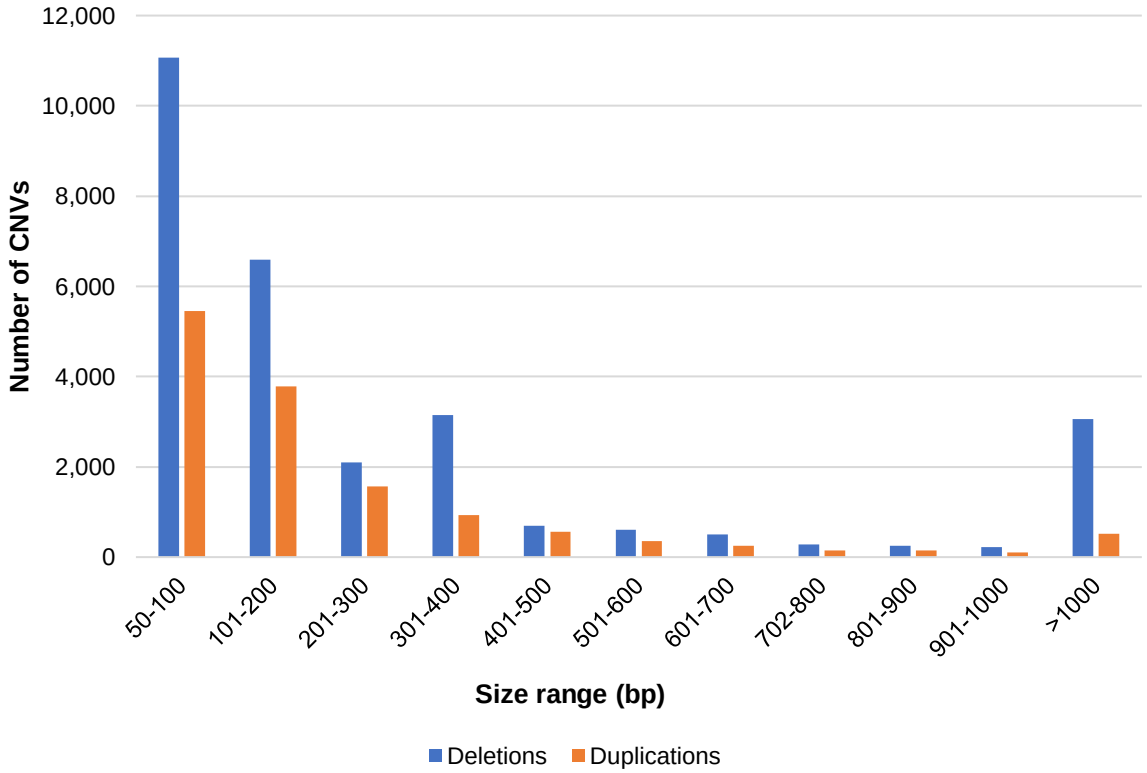


Figure 3.5: Size distribution of deletions (blue) and duplications (orange) across the genome.

The most severe consequence of each CNV was predicted by Ensembl's VEP and can be seen in *Figure 3.6* below. Intronic variants were the most common (46.7%), followed closely by intergenic variants (39.9%). Upstream and downstream variants accounted for 4.9% and 4.3% of CNVs respectively. CNVs affecting various sites important for splicing were also identified, with 250 (0.6%) splice acceptor variants and 150 (0.4%) spice donor variants identified. Additionally, several CNVs affected the untranslated regions (UTRs), with 3' UTR variants and 5' UTR variants accounting for 0.5% and 0.1% of CNVs. More severe consequences were observed at lower frequencies, including: transcript ablation (0.34%); coding sequence variants (0.27%); in-frame deletions (0.2%); transcript amplification (0.04%) and frame shift variants (0.04%). In total, 12 570 genes across the genome were affected by CNVs.

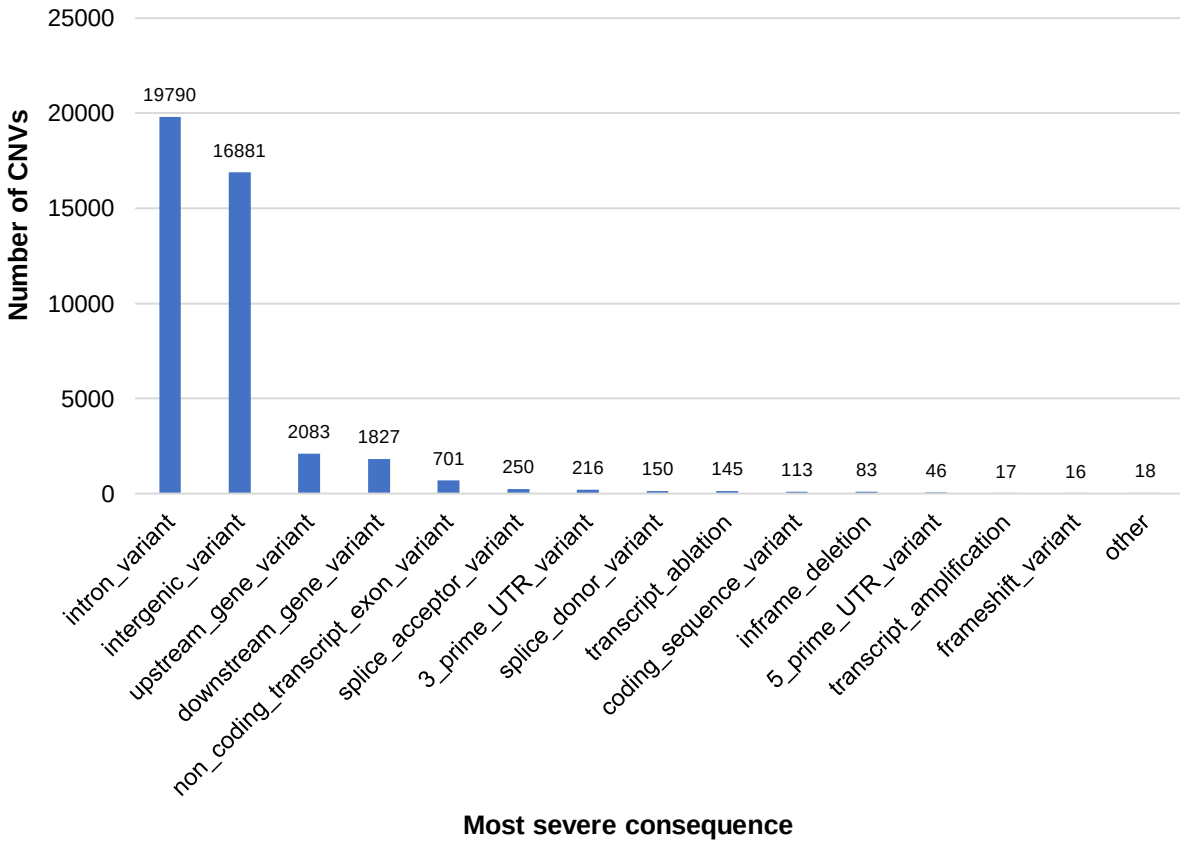


Figure 3.6: Most severe consequence of CNVs as predicted by Ensembl's VEP. "Other" includes splice region variants (n = 11); mature miRNA variants (n = 4); stop losses (n = 2) and start losses (n = 1).

3.3.4 Comparison to short-read CNVs

A comparison between the CNVs identified by the short- and long-read sequencing approaches was performed in order to validate the CNVs identified using a short-read approach, as well as to assess whether short-read sequencing is sufficient for accurate CNV calling or whether long-read sequencing is required. While the focus of this study is on ADME genes, CNVs across the whole genome were assessed in this chapter.

The long-read approach identified more CNVs from far fewer samples compared to the short-read approach (*Table 3.4*). A total of 42 336 CNVs were called from the long-read sequences of only eight individuals (*Table 3.4*). In comparison, Genome STRiP, Manta and SURVIVOR called 15 658, 53 986 and 30 798 CNVs respectively from just under 1 000 short-read sequences (*Table 3.4*). Additionally, the long-read sequences had an average of 16 493 CNVs per sample, whereas the Genome STRiP, Manta and SURVIVOR datasets had an average of only 1 788, 10 554 and 4 144 CNVs per sample respectively (*Table 3.4*). Lastly, the percentage of novel CNVs detected by the long-read approach was again higher (42.9%) when compared to the Genome STRiP, Manta and SURVIVOR CNVs of which 34%, 36.3% and 32% were classified as novel (*Table 3.4*).

Table 3.4: Comparison between the short-read CNVs (Genome STRiP, Manta and SURVIVOR dataset) and long-read CNVs (pbsv) in terms of the number of CNVs detected, average number of CNVs detected per sample and the percentage of novel CNVs detected.

	GenomeSTRiP (n = 919)	**Manta (n = 953)**	**SURVIVOR (n = 953)**	**pbsv (n=8)**
Number of CNVs detected	15 658	53 986	30 798	42 336
Average number of CNVs detected/sample	1 788	10 544	4 144	16 493
Percentage of novel CNVs detected	34%	36.3%	32%	42.9%

CNVs called from the eight individuals for which we had long-read sequencing data were extracted from the short-read datasets and the CNVs were compared to the long-read dataset to assess the overlap (*Figure 3.7*). When the CNVs called by Genome STRiP were compared to those called from the long-read data from the same eight individuals (based on 50% reciprocal overlap), 52.5% (n = 1 816) of the Genome STRiP CNVs were concordant with long-read CNVs; however, only 4.3% of the long-read CNVs were concordant with the Genome STRiP CNVs (*Figure 3.7A*). Given that the size range of CNVs detected by Genome STRiP (1–100 kb) and pbsv (50bp–100 kb) differed, we compared only the CNVs within the size range detected by both tools (1–100 kb). A total of 3 577 CNVs called from the long-read data were between 1 kb and 100 kb in size, and the concordance between the long-read CNVs and the GenomeSTRiP CNVs increased from 4.3% to 50.8%.

When the CNVs called by Manta were compared to those called from the long-read data from the same eight individuals (based on 50% reciprocal overlap), 56.9% (n = 12 001) of the Manta CNVs were concordant with the long-read CNVs, and just over a quarter (28.4%) of the long-read CNVs were concordant with the Manta CNVs (*Figure 3.7B*).

Lastly, when the CNVs in the SURVIVOR dataset were compared to those called from the long-read data from the same eight individuals (based on 50% reciprocal overlap), 56.1% (n = 7 340) of the SURVIVOR CNVs were concordant with the long-read CNVs, and 17.3% of the long-read CNVs were concordant with the SURVIVOR dataset (*Figure 3.7C*).

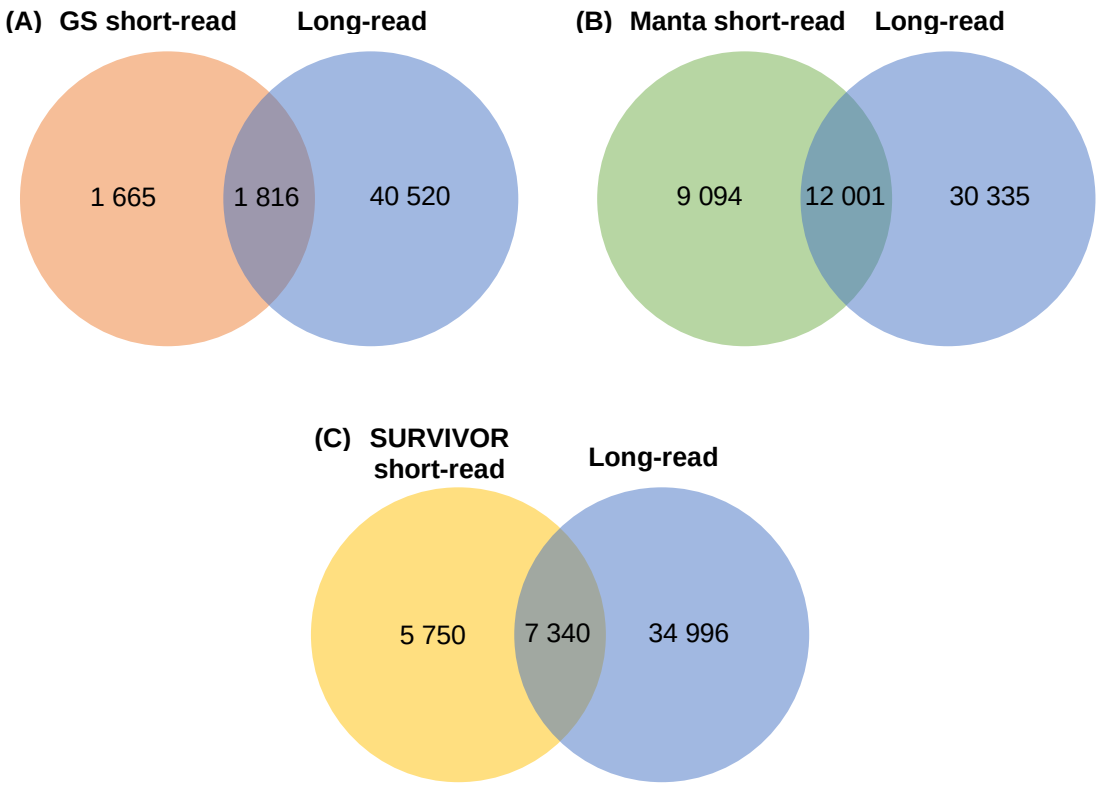


Figure 3.7: Overlap between CNVs called from short-read data (A) Genome STRiP; (B) Manta; (C) SURVIVOR and the long-read sequencing data obtained from the same 8 individuals (based on 50% reciprocal overlap).

3.3.5 CNVs in ADME genes

Of the 42 336 CNVs identified across the genome, only 236 affected ADME genes. In total, 101 ADME genes were affected by these CNVs, including 21 core ADME genes and 80 extended ADME genes. A list of these genes can be seen in *Table 3.5* below.

Table 3.5: List of core (n = 21) and extended (n = 80) ADME genes affected by CNVs called from the long-read data.

Gene	**Core/Extended**
ABCB1	Core
ABCC2	Core
ABCG2	Core
CYP1A1	Core
CYP2C8	Core
CYP2E1	Core
DPYD	Core
EPHX1	Core
FMO2	Core
FMO3	Core
GSR	Core
GSTCD	Core
GSTM4	Core
HNF4A	Core
HNMT	Core
HSD17B11	Core
HSD17B14	Core
SLC15A2	Core
SLC22A1	Core

SLC22A2	Core
UGT2B7	Core
ABCA1	Extended
ABCB11	Extended
ABCB5	Extended
ABCB8	Extended
ABCC3	Extended
ABCC4	Extended
ABCC5	Extended
ABCC9	Extended
ABCG1	Extended
ADH1B	Extended
ADH4	Extended
ALDH1A2	Extended
ALDH1A3	Extended
ALDH3A1	Extended
ALDH3A2	Extended
ALDH3B2	Extended
ALDH6A1	Extended
ALDH7A1	Extended
CBR3	Extended
CDA	Extended
CHST2	Extended
CHST5	Extended
CHST8	Extended
CHST9	Extended
CYB5R3	Extended
CYP19A1	Extended
CYP2A7	Extended
CYP2S1	Extended
CYP39A1	Extended
CYP3A43	Extended
CYP4F11	Extended
CYP4F12	Extended
CYP4F2	Extended
CYP4F3	Extended
CYP4F8	Extended
CYP7B1	Extended
DHRS12	Extended
DHRS4L2	Extended
DPEP1	Extended

NOS1	Extended
NOS3	Extended
PDE3A	Extended
PDE3B	Extended
PON1	Extended
POR	Extended
PPARD	Extended
RXRA	Extended
SLC13A3	Extended
SLC19A1	Extended
SLC22A11	Extended
SLC22A12	Extended
SLC22A16	Extended
SLC22A18	Extended
SLC22A18AS	Extended
SLC27A1	Extended
SLC28A1	Extended
SLC28A2	Extended
SLC2A5	Extended
SLC7A5	Extended
SLCO1C1	Extended
SLCO2A1	Extended
SLCO3A1	Extended
SLCO4A1	Extended
SOD2	Extended
SULF1	Extended
SULT1A2	Extended
SULT1B1	Extended
SULT2B1	Extended
SULT4A1	Extended
UGT1A10	Extended
UGT1A3	Extended
UGT1A4	Extended
UGT1A5	Extended
UGT1A6	Extended
UGT1A7	Extended
UGT1A8	Extended
UGT1A9	Extended
UGT2A1	Extended
UGT2B28	Extended
XDH	Extended

3.4 Discussion

While short-read sequencing approaches are well suited for the detection of single nucleotide variants (SNVs) and small insertions and deletions (INDELs), they do not perform as well for the detection of SVs, including CNVs (Mahmoud *et* al., 2019). The development of long-read sequencing in recent years has led to several improvements in SV calling, with long-read methods for SV calling performing better when compared to short-read methods (Mahmoud *et al*., 2019). In this study, data from eight long-read whole genome sequences (WGSs) was generated with the aim of validating the CNVs called from short-read sequence data, as well as to gain a deeper understanding of the CNV landscape due to the benefits that this approach provided.

In total 42 336 CNVs were called from the long-read sequences, the majority (67.4%) of which were deletions. The relationship between the number of CNVs identified on a chromosome and the size of the chromosome was shown to be non-linear. For example, chromosome 19 had a greater proportion of CNVs when scaled by chromosome length compared to all of the larger chromosomes. This has been seen in previous studies and can be explained by differences in the genomic structure of the chromosomes (De Smith et al., 2009). For example, chromosome 19 contains a large proportion of segmental duplications which may result in an increased presence of CNVs on this chromosome compared to other larger chromosomes (Grimwood et al., 2004; De Smith et al., 2009). The majority of CNVs were small (< 1000 bp), with only 8.4% of CNVs being greater than 1 kb in length. This may be due to the fact that long-read approaches call CNVs predominantly based on sequencing information, rather than read depth (Zhao *et al.*, 2021). The advantage of a depth-based approach is the ability to detect large CNVs (Glusman *et al.*, 2015). Therefore, the lack of depth-based approaches for long-read sequencing data limits the size of CNVs that can accurately be detected (Zhao *et al.*, 2021).

In terms of functional impact, intragenic CNVs (60.1%) were more common than intergenic CNVs (39.9%). This is contradictory to other studies that have shown that intergenic regions are more enriched for CNVs compared to intragenic regions due to negative selection of CNVs within genes (Rigau *et al.*, 2019). However, this may be explained by the fact that intergenic CNVs are on average greater in length compared to intragenic CNVs (Zhang *et al.*, 2021) and as the long-read approach was able to

detect mostly small CNVs (< 1 000bp), the large CNVs that occur more frequently in intergenic regions may not have been detected. Of the intragenic CNVs, intronic variants were the most common (46.7%), and variants resulting in more deleterious predicted consequences, such as transcript ablation and amplification, coding sequence variants and frame shift variants, were observed at very low frequencies (< 1%). This may be due to negative selection of deleterious variants.

When comparing the CNVs called across the eight individuals, ~7% of the CNVs were observed across all samples. The identification of these CNVs across all of the samples is strong evidence of these CNVs being at a very high frequency or fixed in southern African populations. Almost half of the CNVs were unique to an individual or seen in only two individuals, and on average each sample contained 1 813 unique CNVs. When comparing the data from the two individuals of each self-identified ethnolinguistic group, they only shared a maximum of 40% of the same CNVs. Even though we only had data from two individuals representative of an ethnolinguistic group, these results speak to the diversity of CNVs between different individuals, even those originating from the same country and self-identified ethnolinguistic group. Future sequencing of more individuals will likely lead to the identification of additional CNVs as was seen by Ebert and colleagues, where the number of SVs increased as the number of sequences increased, and this increase was even more significant when samples from individuals of African descent were added (Ebert *et al.*, 2021).

Recent studies have shown that SV calling from long-read sequencing data increases the number of SVs detected. A study performed by Ebert and colleagues used PacBio long-read sequencing, in combination with Strand-seq, on 32 samples and a total of 107 126 SVs were identified (Ebert *et al.*, 2021). This more than tripled the rate of discovery when they compared these results to short-read SV analysis on the same samples (Ebert *et al.*, 2021). The number of SVs identified per sample has also been increased by the use of long-read sequencing, with Ebert and colleagues identifying 23 000–28 000 SVs per sample (Ebert *et al.*, 2021). We found similar findings in this study, with the long-read sequencing approach identifying more CNVs from far fewer samples, compared to the short-read sequencing approach. We identified 42 336 CNVs from only eight long-read sequences compared to between ~15 000 and ~50 000 CNVs using different calling approaches, from just under 1 000 short-read sequences. We also observed an increase in the number of CNVs called per sample,

with average of 16 493 CNVs per sample for the long-read sequencing data compared to between ~1 700 and ~ 10 000 CNVs per sample for the short-read sequencing approaches.

The percentage of novel CNVs detected by the long-read approach was higher (42.9%) when compared to the CNVs called by Genome STRiP, Manta and SURVIVOR, of which 34%, 36.3% and 32% respectively were classified as novel. The higher proportion of novel CNVs identified by the long-read approach may be attributed to the fact that a large proportion (39%) of the CNVs called using this approach were very small (< 100 bp). This size range (50–100 bp) of CNVs has not been as well studied and is underrepresented in databases such as the DGV.

When focusing on the eight samples with matching short- and long-read sequence data, approximately half of the CNVs detected from the short-read approaches were concordant with long-read data. However, the percentage of long-read CNVs that were concordant with short-read CNVs ranged from only 4.3% to 28.4% depending on the CNV calling tool used. The concordance between the long-read CNVs and the Genome STRiP CNVs did increase from 4.3% to 50.8% when comparing CNVs that fell within the size range detected by both tools. Ebert and colleagues (note that they assessed all SVs, while we only assessed deletions and duplications) found that 77.4% of SVs detected from the short-read sequences were concordant with long-read variants. This is higher than what we observed in our current study. They also observed that only 29.6% of long-read SVs were observed in the short-read dataset, which is similar to results obtained from our study. Lastly, Ebert and colleagues found that the greatest gains in sensitivity from the long-read approach were observed among smaller SVs, which was replicated in this study, with ~40% of the CNVs ranging in size from 50 to 100 bp in length, and over 80% of the CNVs being smaller than 500 bp in length (Ebert *et al.*, 2021).

The results of this study show that long-read sequencing is required for more accurate and comprehensive CNV calling, particularly for small CNVs. However, this does not mean that it should completely replace the use of short-read sequencing for CNV calling. Ebert and colleagues found that for larger variants, a short-read sequencing approach works fairly well; however, for smaller variants (50-300 bp) there is almost an order of magnitude difference between long-read and short-read call sets (Ebert *et*

al., 2021). In the future, NGS-based CNV calling will most probably rely on a combination of short- and long-read sequencing approaches (Gabrielaite *et al.*, 2021). However, cost and throughput are the main factors that limit the number of genomes being sequenced using a long-read approach (Huddleston *et al.*, 2017). Unfortunately the current cost of long-read sequencing is significantly more than that of short-read sequencing (Zhao *et al.*, 2021). Additionally, the long-read sequencing approaches have a comparatively lower throughput compared to short-read approaches which can be used on hundreds of thousands of samples in a single study, making long-read sequencing currently impractical for large-scale population studies (Zhao *et al.*, 2021). A potential solution to cost and throughput issues mentioned above, is calling CNVs from genotyping data that has previously been generated for genome wide association studies (GWAS). In the following chapter, this approach will be explored.

To the best of our knowledge, we have generated one of the first sets of long-read whole-genome sequences from African individuals. This approach allowed for the detection of over 40 000 CNVs across the genome, which not only validated many of the CNVs detected from the short-read approach but also allowed us to gain a deeper understanding of the CNV landscape in southern African populations. Using a long-read sequencing approach, a greater percentage of novel CNVs were identified, and many very small CNVs, that have not been well studied previously, were identified. Additionally, we identified a large number of CNVs that were common across all eight study participants, which suggests that these CNVs are present at very high frequencies in southern African populations. However, there were also a large proportion of variants that were unique to each individual suggesting that there is a large amount of inter-individual variation, even between individuals of the same self-identified ethnolinguistic group. Overall, the use of a long-read sequencing approach has provided invaluable insights into the CNV landscape, and sequencing of additional samples will further improve our knowledge thereof.

3.5 Reference list

Abel, H.J., Larson, D.E., Regier, A.A., *et al.* (2020) Mapping and characterization of structural variation in 17,795 human genomes. *Nature*. 583(7814), pp. 83–89. doi:10.1038/s41586-020-2371-0.

Collins, R.L., Brand, H., Karczewski, K.J., *et al.* (2020) A structural variation reference for medical and population genetics. *Nature*. 581(7809), pp. 444–451. doi:10.1038/s41586-020-2287-8.

Ebert, P., Audano, P.A., Zhu, Q., *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science.* 372(6537), pp. 1–33. doi:10.1126/science.abf7117.

Gabrielaite, M., Torp, M.H., Rasmussen, M.S., *et al.* (2021) A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers.* 13(24), pp. 1–21. doi:2021.04.30.442110. doi:10.1101/2021.04.30.442110.

Glusman, G., Severson, A., Dhankani, V., *et al.* (2015) Identification of copy number variants in whole-genome data using Reference Coverage Profiles. *Frontiers in Genetics*. 6(45), pp. 1–13. doi: 10.3389/fgene.2015.00045.

Grimwood, J., Gordon, L.A., Olsen, A., et al. (2004) The DNA sequence and biology of human chromosome 19. *Nature*. 428(6982), pp. 529–535. doi:10.1038/nature02399.

Hon, T., Mars, K., Young, G., *et al.* (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*. 7(1), pp. 399. doi:10.1038/s41597-020-00743-4.

Howe, K.L., Achuthan, P., Allen, James, *et al.* (2021) Ensembl 2021. *Nucleic Acids Research*. 49, pp. 884–891. doi:10.1093/nar/gkaa942.

Huddleston, J., Chaisson, M.J.P., Steinberg, K.M., *et al.* (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*. 27(5), pp. 677–685. doi:10.1101/gr.214007.116.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., *et al.* (2019) Structural variant calling: the long and the short of it. *Genome Biology*. 20(1), pp. 246. doi:10.1186/s13059-019-1828-7.

PacBio (2020) *Sequencing 101: The Evolution of DNA Sequencing Tools*. Available at: https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/ (Accessed: 01/10/2020).

Ramsay, M., Crowther, N., Tambo, E., *et al.* (2016) H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Global health, epidemiology and genomics*. 1(20), pp. 1–13. doi:10.1017/gheg.2016.17.

Rigau, M., Juan, D., Valencia, A., *et al.* (2019) Intronic CNVs and gene expression variation in human populations. *PLoS Genetics*. 15(1), pp. 1–23. doi:10.1371/journal.pgen.1007902.

Sedlazeck, F.J., Lee, H., Darby, C.A., *et al.* (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*. 19(6), pp. 329–346. doi:10.1038/s41576-018-0003-4.

de Smith, A.J., Walters, R.G., Froguel, P., et al. (2009) Human genes involved in copy number variation: Mechanisms of origin, functional effects and implications for disease. *Cytogenetic and Genome Research*. 123(4), pp. 17–26. doi:10.1159/000184688.

del Tredici, A.L., Malhotra, A., Dedek, M., *et al.* (2018) Frequency of CYP2D6 Alleles Including Structural Variants in the United States. *Frontiers in Pharmacology*. 9, pp. 305. doi:10.3389/fphar.2018.00305.

Wenger, A.M., Peluso, P., Rowell, W.J., *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*. 37(10), pp. 1155–1162. doi:10.1038/s41587-019-0217-9.

Zhang, L., Shi, J., Ouyang, J., *et al.* (2021) X-CNV: genome-wide prediction of the pathogenicity of copy number variations. *Genome Medicine*. 13(1), pp. 132. doi:10.1186/s13073-021-00945-4.

Zhao, X., Collins, R.L., Lee, W.-P., *et al.* (2021) Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *The American Journal of Human Genetics*. 108(5), pp. 919–928. doi:https://doi.org/10.1016/j.ajhg.2021.03.014.

CHAPTER 4: AN ASSESSMENT OF THE UTILITY OF THE INFINIUM H3AFRICA CONSORTIUM ARRAY V2 TO CALL CNVS

4. CHAPTER 4: AN ASSESSMENT OF THE UTILITY OF THE INFINIUM H3AFRICA CONSORTIUM ARRAY V2 TO CALL CNVS

4.1 Introduction

A variety of technologies can be used to detect copy number variants (CNVs), including next-generation sequencing (NGS) approaches such as short- and long-read whole genome sequencing, as discussed in previous chapters, as well as genome-wide single nucleotide polymorphism (SNP) arrays. The high-throughput, high density genotyping arrays that are used in genome-wide association studies (GWAS) have been increasingly used for the detection of CNVs (Carter, 2007; Lin, Naj and Wang, 2013). Although originally developed in order to determine the bases present at certain SNPs, the signal intensity information derived from these arrays can be used to detect CNVs (Carter, 2007). As millions of individuals have already been genotyped on such arrays, a wealth of unexploited information remains available in this data which could be reanalysed for CNV detection (Macé *et al.*, 2016).

A limitation of using these arrays for CNV detection is the lower level of reliability that is achieved as they were not initially designed for CNV detection (Macé *et al.*, 2016). However, some of the newer arrays have included several CNV probes (Macé *et al.*, 2016). In spite of this limitation, several tools have been developed for the detection of CNVs from genotyping arrays, with tools either being based on individual-wise analysis or joint analysis of multiple individuals (Zhang *et al.*, 2019). In this study we made use of PennCNV (Wang *et al.*, 2007) which is currently the most widely used software for CNV detection from Illumina arrays (Macé *et al.*, 2016). PennCNV implements a hidden Markov model (HMM) that integrates multiple sources of information generated by genotyping arrays to call CNVs from individual samples (Wang *et al.*, 2007). The log R ratio (LRR) and B-allele frequency (BAF) information from each probe is used to determine the copy number (CN) (Wang *et al.*, 2007; Macé *et al.*, 2016), and additional information, such as population frequencies at each locus and linkage-disequilibrium (LD) structure between adjacent loci, is used to improve the quality of the CNV calls (Wang *et al.*, 2007; Zhang *et al.*, 2019). In general, individual-wise approaches are successful in detecting large, rare CNVs spanning at least ten probes on a SNP array (Zhang *et al.*, 2019). However, as these methods are reliant on the assumption that allele intensities are properly normalized, this type of analysis can be error-prone and

strict quality control of the CNV calls is required (Macé *et al.*, 2016; Zhang *et al.*, 2019). Moreover, comparing CNVs between different individuals can be difficult as the individual-level CNV calls are often uniform across individuals (Zhang *et al.*, 2019).

The Human Heredity and Health in Africa (H3Africa) Consortium Array is a powerful Illumina array designed using novel African content, making it the most effective array to date for the study of African populations. Relevant to this study, the array includes 24 168 probes within core and extended ADME genes. However, this is a general purpose array, and is not specifically designed for CNV detection. Therefore, the aim of this chapter is to determine which CNVs can be called from the Infinium H3Africa Consortium Array v2, thus allowing for the assessment of the utility of the array to call CNV's.

An alternative approach to detect CNVs in tagging. CNVs that have previously been detected, from sequencing approaches for example, can be tested for associations with SNP haplotypes. In this way a SNP haplotype that tags a CNV can be used to predict the presence or absence of a CNV, as well as predict the likely CN of the CNV, without the need to perform any CNV calling (Nyangiri *et al.*, 2020). Siwada (2020)[1] investigated this approach for CNVs in ADME genes and found that less than a quarter of the ADME CNVs identified were tagged and thus could be predicted using the Infinium H3Africa Consortium Array v2 (Siwada, 2020). As the yield was relatively small, in combination with the fact that the majority of CNVs in core ADME genes were not tagged, this was not investigated further in this study.

[1] Honours student in the Division of Human Genetics, under my supervision.

4.2 Materials and Methods

4.2.1 Study participants

The genotyping data used in this analysis was obtained from the H3Africa AWI-Gen project (Ramsay *et al.*, 2016). All individuals were genotyped on Illumina's Infinium H3Africa Consortium Array v2. The genotyping data from 10 857 individuals was used in this study, consisting of 4 848 participants from South Africa, 2 079 participants from Burkina Faso, 1 988 participants from Ghana and 1 942 participants from Kenya. A map showing the locations of the participants can be seen in *Figure 4.1* below.

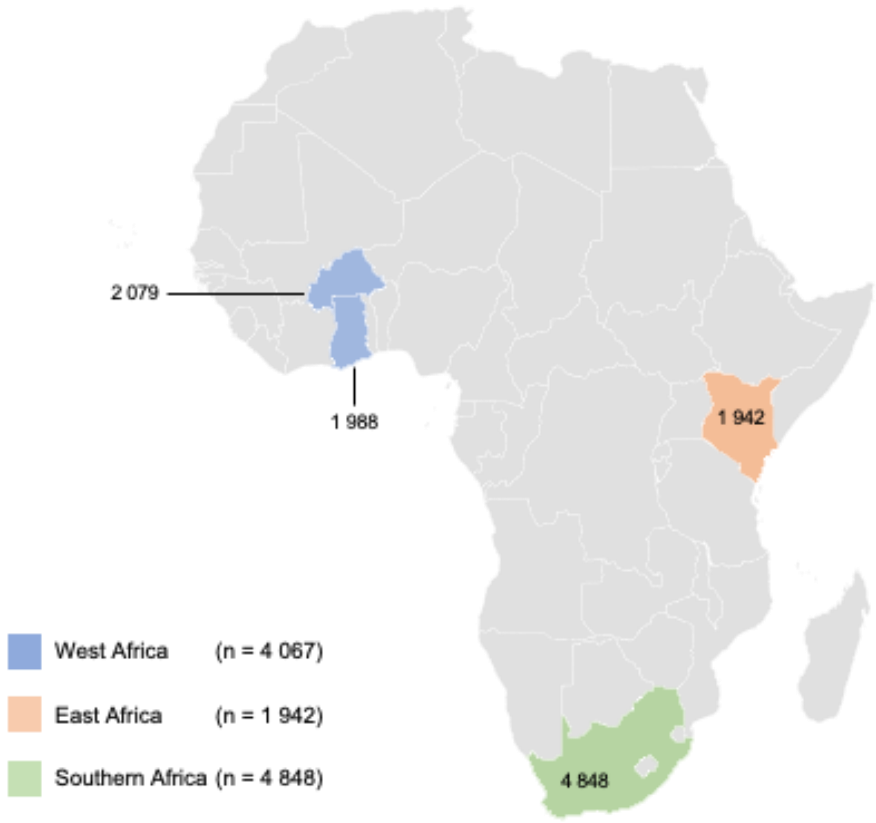


Figure 4.1: Map showing the locations of participants in the H3Africa AWI-Gen study.

4.2.2 Data analysis

All of the scripts and files that were included in the PennCNV package (v 1.0.5) (Wang *et al.*, 2007) and used in the following analysis can be found in the following GitHub repository: https://github.com/WGLab/PennCNV.

PennCNV input requirements

Illumina provides a single report containing the signal intensity values for all samples; however, PennCNV requires separate reports for each sample containing the LRR and BAF values. PennCNV provides a script (split_illumina_report.pl) which splits the Illumina report into files containing each sample individually. As the Illumina data used in this study consisted of multiple batches and the reports were split into a large number of smaller files, a Nextflow script was used to split the reports simultaneously. This script as well as the configuration file can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%204%20scripts/split_illumina_report.nf and https://github.com/lauracottino/PhD_Thesis/blob/main/Chapter%204%20scripts/split_illumina_report.config). In total, 10 857 signal intensity files were generated. The second requirement of PennCNV is a Population Frequency of the B allele (PFB) file. For any non-European data, such as the data used in this study, PennCNV requires that a PFB file be generated from the signal intensity files. PennCNV includes a script to generate the PFB file (compile_pfb.pl) and requires a file with the chromosome co-ordinates of the all the SNPs on the array as input.

CNV calling with PennCNV

CNV calling was performed using PennCNV's detect_cnv.pl script using default parameters. The script required a list of the individual sample signal intensity files, the PFB file created above as well as an hmm file. PennCNV provides such a file (hhall.hmm) which is said to be suitable for all Illumina arrays. The 'test' option was used to test the HMM model to identify CNVs. The script outputs CNV calls from each sample in a single VCF file. Only CNVs on the 22 autosomes were assessed in this study due to the inaccuracy of CNV calling on the sex chromosomes. CNVs called from the autosomes are more reliable than those called from sex chromosomes in terms of size and breakpoint accuracy. Careful evaluation and additional validation would be require for CNVs detected on the sex chromosomes (Yao *et al*., 2019).

Quality control of PennCNV data

CNV calling tools routinely split large CNVs into smaller adjacent CNVs. PennCNV provides a script (clean_cnv.pl) to merge adjacent CNV calls. During this process 1 270 180 raw CNVs were merged into 1 219 090 raw CNVs.

Sample level filtering was performed using PennCNV's filter_cnv.pl script. This script computed the following statistics for each sample: (1) LRR SD (standard deviation of LRRs); (2) BAF drift (measuring departure of the BAF from the expected values); (3) WF (waviness factor, the amount of dispersion in signal intensity, which is a good indicator for DNA quality after accounting for GC content); and (4) NumCNV (number of called CNVs). This data was then used to exclude samples based on default exclusion criteria: LRR SD > 0.3, BAF drift > 0.01 and WF > 0.05, whereby samples that meet any of the above three criteria were excluded from further analysis. Additionally samples with 115 or more CNVs per sample were considered outliers and removed. Through this filtering process, the number of raw CNVs was reduced from 1 219 090 to 816 956.

The number of SNPs spanned by each CNV was specified in the PennCNV output file (numsnp field). This was assessed in order to perform CNV level filtering. Generally, a minimum of three SNPs is used; however, Lin and colleagues suggests that a minimum of 10 SNPs produces results with reasonable sensitivity (Lin, Naj and Wang, 2013). Filtering of CNVs spanning 10 or more SNPs was performed using the filter_cnv.pl script with the -numsnp argument. Following filtering, a total of 396 689 raw CNVs remained.

BEDTools (v2.27.1) was used to merge CNVs across the remaining samples. During this step 396 689 raw CNVs were reduced to 7 370 distinct CNVs.

CNV Annotation

Basic Linux commands were used for data analysis. CNVs identified from the array were compared to those called from the short- and long-read sequences using BEDTools intersect (v2.27.1) with 50% reciprocal overlap. The CNVs called from the array have build 37 co-ordinates, therefore in order to be able to compare these results to the short- and long-read CNVs, which have build 38 co-ordinates, lift over of the

array CNV co-ordinates was required. Lift over was performed on UCSC's genome browser (https://genome.ucsc.edu/cgi-bin/hgLiftOver). As the output of PennCNV is a text file and not a VCF, the text file was first converted to a BED file which in turn was converted to a VCF using SURVIVOR's (v1.0.7; https://github.com/fritzsedlazeck/SURVIVOR/) bedtovcf function. Ensembl's Variant Effect Predictor (VEP) (v104) (Howe *et al.*, 2021) was used to determine which genes were overlapped by CNVs as well as to assess the consequences of the CNVs. All graphs were plotted using Microsoft Excel (v16.54).

4.3 Results

4.3.1 Summary of array CNVs

A total of 7 312 CNVs were identified from the genotyping array data of just over 10 000 individuals. A file containing these CNVs can be seen in the GitHub repository (https://github.com/lauracottino/PhD_Thesis/blob/main/Supplementary%20Files/Array_CNVs.xlsx). *Figure 4.2* below shows the distribution of the array CNVs across the chromosomes. The vast majority of CNVs (n = 6 888; 94.2%) were rare, with a frequency of less than 1%. Only 4% of CNVs (n = 293) were observed at a low frequency (1–5%) and even fewer CNVs were common, with a frequency of greater than 5% (n = 131; 1.8%). Individuals harboured between 11 and 83 CNVs, with an average of 33 CNVs per individual. Of the 7 312 CNVs, 133 affected ADME genes. In total, 140 ADME genes were affected by these CNVs, including 21 core ADME genes and 119 extended ADME genes. A list of these genes can be seen in *Table 4.1* below.

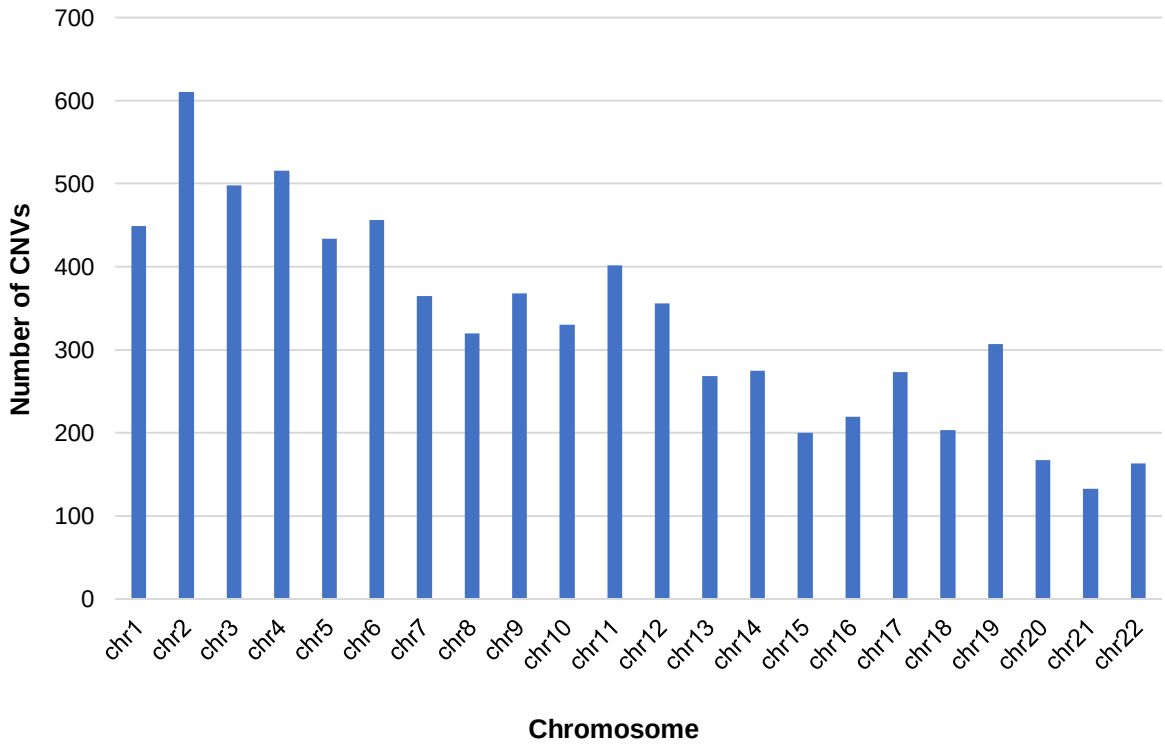


Figure 4.2: Distribution of CNVs across the chromosomes.

Table 4.1: List of core (n = 21) and extended (n = 119) ADME affected by CNVs called from the Infinium H3Africa Consortium Array v2.

ADME gene	**Core/Extended**
ABCB1	Core
ABCC2	Core
CYP1A1	Core
CYP1A2	Core
CYP2A6	Core
CYP2B6	Core
CYP2C19	Core
CYP2C8	Core
CYP2C9	Core
CYP2D6	Core
CYP2E1	Core
DPYD	Core
GSTM1	Core
NAT1	Core
SLCO1B1	Core
SLCO1B3	Core
SULT1A1	Core
TPMT	Core
UGT2B15	Core
UGT2B17	Core
UGT2B7	Core
ABCA1	Extended
ABCC11	Extended
ABCC13	Extended
ABCC4	Extended
ABCC8	Extended
ABCG1	Extended
ADH1B	Extended
ADH1C	Extended
AHR	Extended
ALDH1A2	Extended
ALDH1A3	Extended
ALDH2	Extended
ALDH3A1	Extended
ALDH3A2	Extended
ALDH3B2	Extended
ALDH4A1	Extended
ALDH7A1	Extended
ARSA	Extended
ATP7B	Extended
CBR3	Extended
CDA	Extended
CES1	Extended
CES2	Extended
CFTR	Extended
CHST11	Extended
CHST12	Extended
CHST5	Extended
CHST6	Extended
CHST8	Extended
CYB5R3	Extended
CYP11A1	Extended
CYP11B1	Extended
CYP11B2	Extended
CYP19A1	Extended
CYP1B1	Extended
CYP21A2	Extended
CYP2A13	Extended
CYP2A7	Extended
CYP2C18	Extended
CYP46A1	Extended
CYP4F11	Extended
CYP4F12	Extended
CYP7B1	Extended
DHRS12	Extended
DHRS4	Extended
DHRS4L1	Extended
DHRS4L2	Extended
DHRS7C	Extended
DPEP1	Extended
GPX1	Extended
GPX4	Extended
GSTA1	Extended

GSTCD	Extended
GSTM2	Extended
GSTM3	Extended
GSTM4	Extended
GSTM5	Extended
GSTT2	Extended
HNF4A	Extended
IAPP	Extended
MAT1A	Extended
METAP1	Extended
MGST1	Extended
NAT2	Extended
PDE3A	Extended
PDE3B	Extended
PNMT	Extended
PON1	Extended
POR	Extended
RXRA	Extended
SLC13A1	Extended
SLC13A2	Extended
SLC15A1	Extended
SLC19A1	Extended
SLC22A10	Extended
SLC22A12	Extended
SLC22A14	Extended
SLC22A18	Extended
SLC22A18AS	Extended
SLC22A9	Extended
SLC2A5	Extended
SLC5A6	Extended
SLC6A6	Extended
SLC7A5	Extended
SLC7A7	Extended
SLCO1A2	Extended
SLCO1C1	Extended
SLCO2B1	Extended
SLCO3A1	Extended
SLCO4A1	Extended
SLCO5A1	Extended
SLCO6A1	Extended
SOD1	Extended
SOD3	Extended
SULF1	Extended
SULT1A2	Extended
SULT1A3	Extended
SULT1B1	Extended
SULT1C2	Extended
SULT1E1	Extended
SULT2A1	Extended
SULT2B1	Extended
SULT4A1	Extended
TAP2	Extended
UGT1A10	Extended
UGT1A3	Extended
UGT1A4	Extended
UGT1A5	Extended
UGT1A6	Extended
UGT1A7	Extended
UGT1A8	Extended
UGT1A9	Extended
UGT2A1	Extended
UGT2B10	Extended
UGT2B11	Extended
UGT2B28	Extended
UGT2B4	Extended
UGT8	Extended
XDH	Extended

The CNVs identified ranged in size from 309 bp to 8.9 Mb, with the majority of CNVs (n = 4 505; 61.6%) being between 10 kb and 100 kb in size (*Figure 4.3*). 19.9% of CNVs (n = 1 455) were between 1–10 kb and 16.7% of CNVs (n = 1 219) were between 100 kb and 1 Mb (*Figure 4.3*). Very few CNVs were less than 1 kb in size (n = 10; 0.1%), and only 1.7% of CNVs (n = 124) were greater than 1 Mb (*Figure 4.3*).

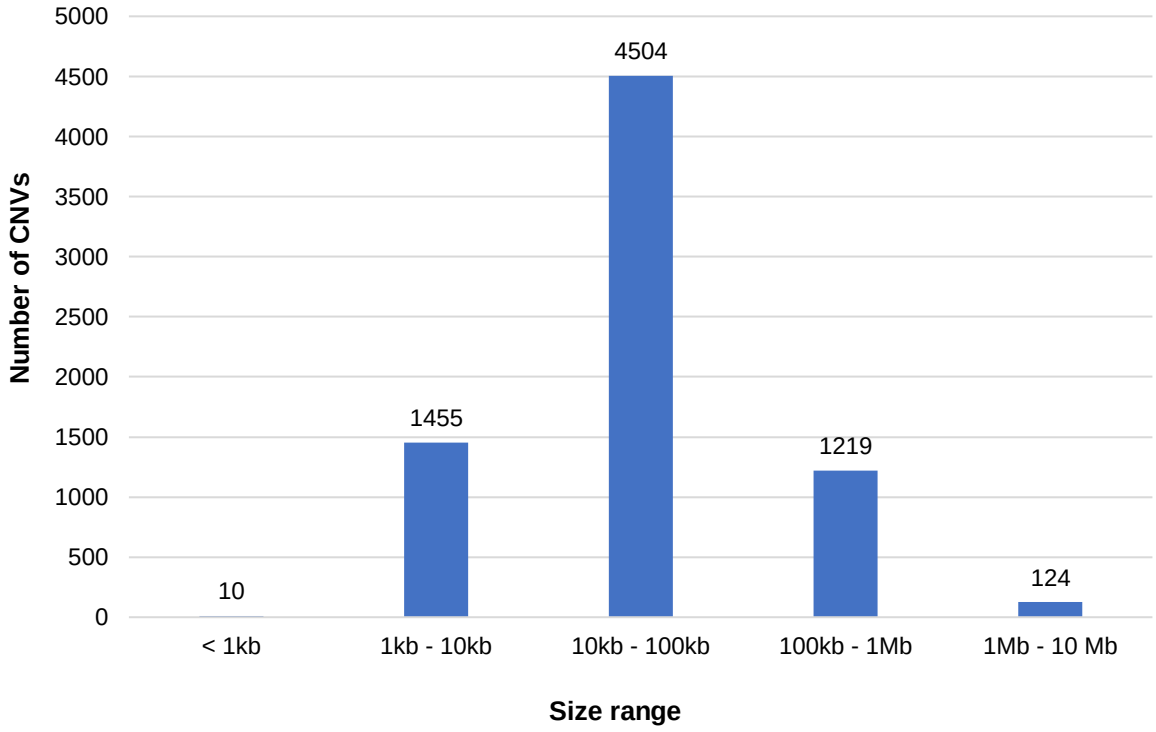


Figure 4.3: Size range of CNVs called from the array data.

The most severe consequences of CNVs called from array data, predicted by Ensembl's VEP, can be seen in *Figure 4.4* below. The most severe consequence for 42.1% (n = 3 078) of the CNVs was transcript ablation or amplification. This was followed by intergenic and intronic variants which accounted for 21.1% (n = 1 546) and 13% (n = 952) of CNVs respectively. Coding sequence variant was the most severe consequence for only 6% of CNVs (n = 439). Upstream and downstream variants accounted for 2.1% (n = 152) and 1.6% (n = 114) of CNVs respectively, while 5' and 3' untranslated region (UTR) variants accounted for 1.6% (n = 119) and 0.6% (n = 47) of CNVs respectively.

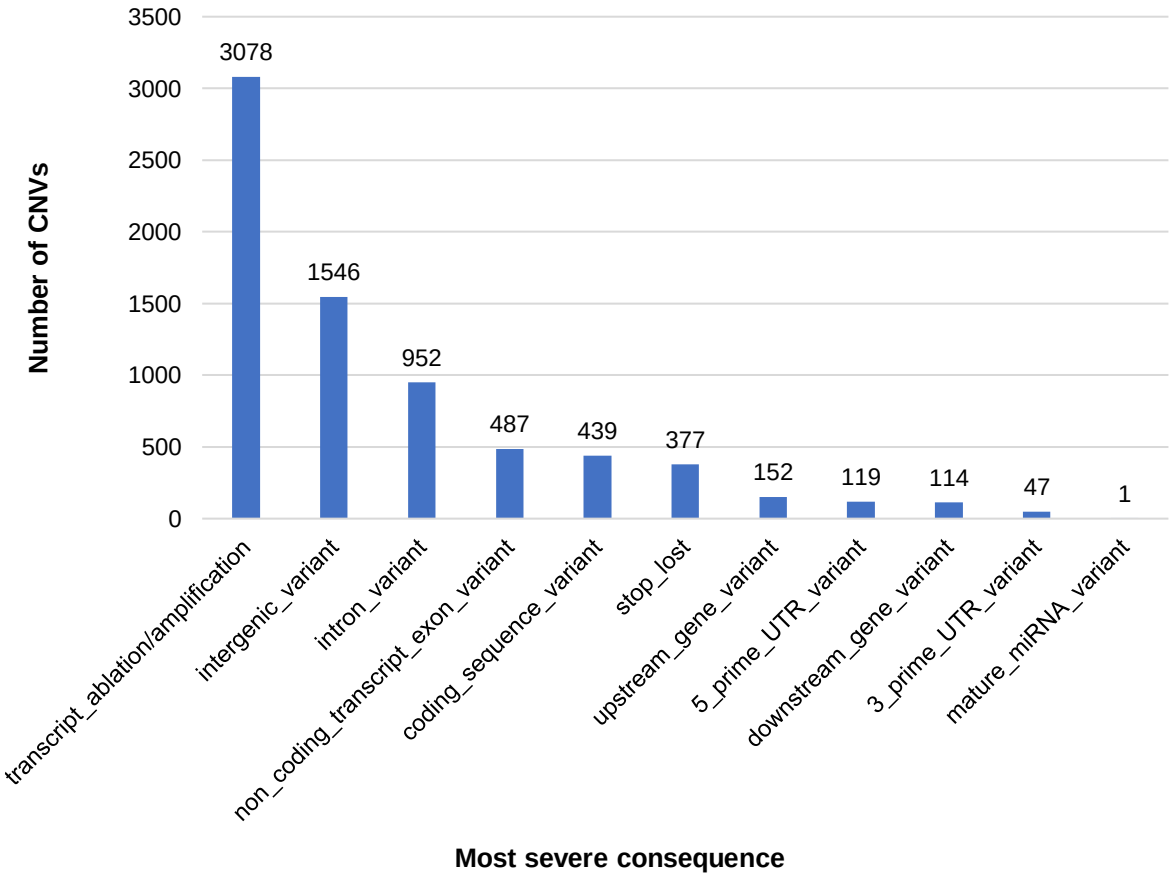


Figure 4.4: Most severe consequence of CNVs as predicted by Ensembl's VEP.

4.3.2 Comparison to short-read data and long-read data

For 100 individuals from South Africa, we had CNV calls from both short-read data and array data. A total of 3 330 CNVs were identified from the genotyping data of the 100 individuals. From the same individuals, the short-read sequencing approach identified more CNVs, with GenomeSTRiP, SURVIVOR and Manta calling 7 703, 21 777 and 33 703 CNVs respectively. The concordance between CNVs called from the array data and short-read data was low. When comparing the CNVs called with Genome STRiP and CNVs called from the array data, only 1.5% of the Genome STRiP CNVs were concordant with the array CNVs, while 3.5% of the array CNVs were concordant with the Genome STRiP CNVs (*Figure 4.5A*). When comparing the CNVs called with SURVIVOR to those called from array data, only 1% of the SURVIVOR CNVs were concordant with the array CNVs, while 6.4% of the array CNVs were concordant with the SURVIVOR CNVs (*Figure 4.5C*). Lastly, when the CNVs called with Manta were compared to the CNVs called from array data, only 0.7% of the Manta CNVs were concordant with the array CNVs, while 7.5% of the array CNVs were concordant with the Manta CNVs, which is more than double the concordance seen between the array CNVs and the CNVs called from GenomeSTRiP data (*Figure 4.5B*).

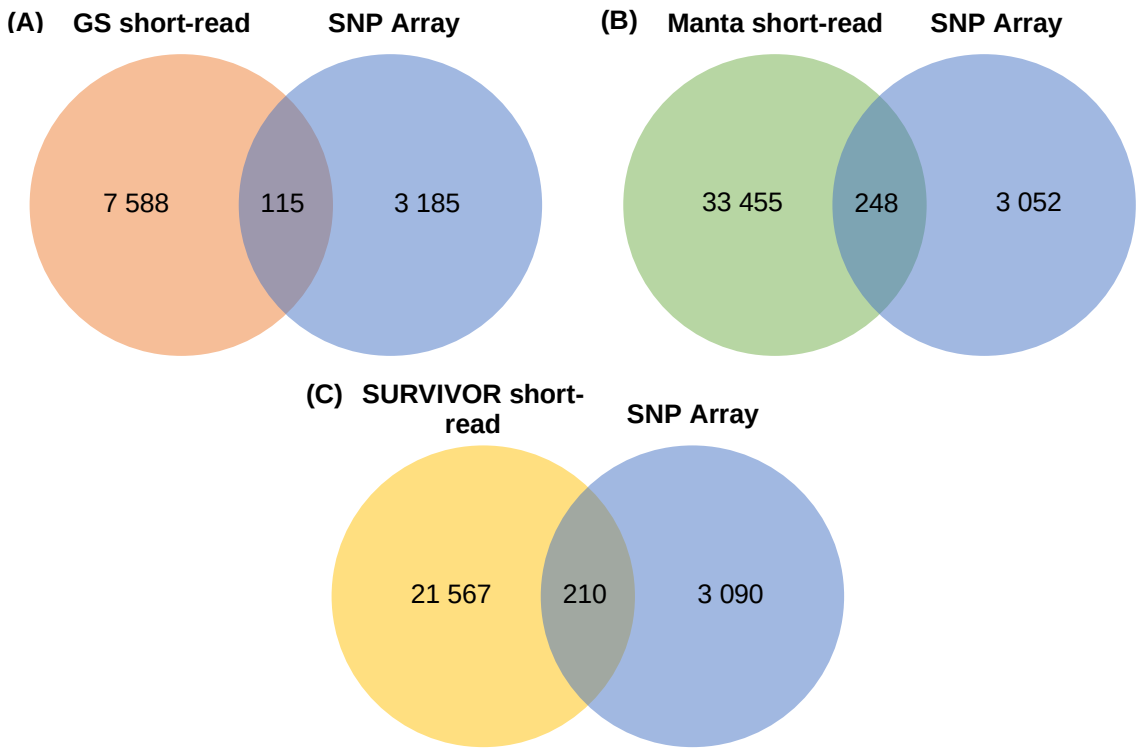


Figure 4.5: Overlap between CNVs called from the array data and short-read sequencing data (A) Genome STRiP; (B) Manta; (C) SURVIVOR from the same 100 South African individuals (based on 50% reciprocal overlap).

The eight samples for which we generated long-read sequencing data had also been genotyped on the Infinium H3Africa Consortium Array v2. The long-read approach identified far more CNVs (n = 42 336) compared to the array-approach which only identified 169 CNVs, when comparing the same eight samples. Interestingly, 53.3% of the CNVs identified from the array data were concordant with the CNV called from long-read data; however, only 0.2% of the long-read CNVs were concordant with the array data (*Figure 4.6*).

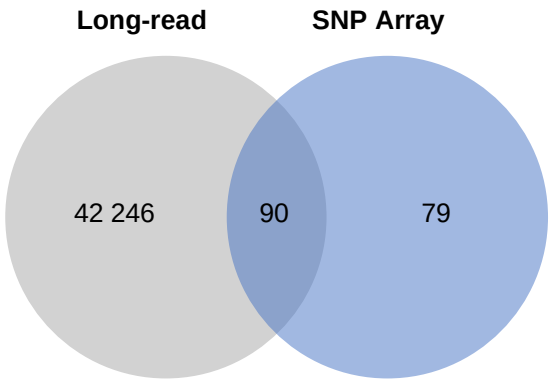


Figure 4.6: Overlap between CNVs called from the array data and long-read sequencing data from the same eight individuals (based on 50% reciprocal overlap).

4.4 Discussion

Various technologies can be used for the detection of CNVs, including short- and long-read whole genome sequencing as well as genotyping arrays (Coutelier *et al.*, 2021). However, while sequencing approaches provide improved CNV calling, they remain prohibitively expensive for large-scale studies (Quick *et al.*, 2020), despite a decline in the cost of sequencing in recent years (Sham and Purcell, 2014). This is particularly relevant for resource-limited regions such as SSA. Genotyping/SNP arrays offer a cost-effective alternative that could potentially allow for large-scale detection of CNVs.

In general, individual-wise approaches, such as PennCNV, are successful in detecting large, rare CNVs (Zhang *et al.*, 2019). This is evident from the results of this study where the vast majority of CNVs (94.2%) were rare, and 80% of CNVs were greater than 10 kb in size. When assessing the most severe consequence for each CNV, the most common consequence was transcript ablation or amplification (42.1%). Partial gene CNVs including intronic variants and coding sequence variants accounted for only 13% and 6% of CNVs respectively. This high proportion of full gene CNVs compared to partial gene CNVs may be due to the fact that most of the CNVs were large and would therefore span the entire gene/s rather than only part of a gene. Additionally, intergenic CNVs were less common (21%) than full gene CNVs (transcript amplification/ablation) which may be due to a bias of probes within genes compared to those in intergenic regions on the array.

A study performed on a Tunisian population, which also made use of PennCNV, identified 751 CNVs from the genotyping data of 102 individuals (Romdhane *et al.*, 2021). The CNVs that were identified by Romdhane and colleagues ranged from 1.02 kb to 3.18 Mb and had an average size of 104 kb (Romdhane *et al.*, 2021). In this study we called CNVs from the genotyping data of over 10 000 individuals and were able to identify 7 312 CNVs. The CNVs that were identified ranged from 309 bp to 8.9 Mb in size with an average size of 104.4 kb, which is comparable to the findings from Romdhane and colleagues (Romdhane *et al.*, 2021). Additionally, we found that the number of CNVs per individual ranged from 11–83 CNVs, with an average of 33 CNVs per individual. This is concordant with the results from Romdhane and colleagues who found that the number of CNVs per individuals ranged from 17–63 CNVs, with an average of 39 CNVs per individual (Romdhane *et al.*, 2021).

It has been shown, that when compared to WGS, genotyping arrays have much lower resolution and therefore detect significantly fewer CNVs (Zhou *et al.*, 2018). This was evident in this study where 3 330 CNVs were identified from the genotyping data of 100 individuals and from the same individuals, the short-read sequencing approach identified ~8 000–34 000 CNVs depending on the CNV calling tool that was used. Unfortunately, the concordance between the array data and short-read data was low with only 0.7–1.5% of the short-read CNVs being concordant with the array CNVs, and 3.5–7.5% of the array CNVs being concordant with the short-read CNVs, depending on the CNV calling tool that was used.

When comparing the CNVs called from the array data and long-read data from the same eight samples, 53.3% of the CNVs identified from the array data were concordant with the CNV called from long-read data. However, the array identified significantly fewer CNVs compared to the long-read approach (169 versus 42 336 CNVs), which is expected as the array is not representative of the entire genome. As such only 0.2% of the long-read CNVs were concordant with the array data. Additionally, the vast majority of CNVs called by the long-read approach were small (< 1000 kb), and only 0.1% of array CNVs were within this size range, making the window of potential overlap extremely small.

While high-coverage whole genome sequencing (WGS) is the current gold standard for population-scale CNV detection (Png *et al.*, 2019), the cost and computational requirements of this approach remain significantly higher than that of arrays (Zhou *et al.*, 2018). Although predominantly used for SNP genotyping and GWAS, the data generated by these arrays can also be used for CNV calling (Carter, 2007). In this study we took advantage of the genotyping data that had already been generated for over 10 000 SSA individuals on the Infinium H3Africa Consortium Array v2 in order to call CNVs. While the results were comparable to other studies, the concordance between the CNVs called from the short-read sequencing data and array data is low, which calls into question the accuracy of the CNV calls. However, the higher concordance obtained between the array data and long-read sequencing data increases our confidence in the array results and thus the arrays utility to accurately call CNVs. However, the array has a much lower resolution, with very few CNVs being called when compared to the short- and long-read WGS data. This limits the utility of the array in generating a comprehensive population-wide CNV call set.

4.5 Reference list

Carter, N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*. 39(7), pp. 16–21. doi:10.1038/ng2028.

Coutelier, M., Holtgrewe, M., Jäger, M., *et al.* (2021) Combining callers improves the detection of copy number variants from whole-genome sequencing. *European Journal of Human Genetics*. 30, pp. 178–186. doi:10.1038/s41431-021-00983-x.

Howe, K.L., Achuthan, P., Allen, James, *et al.* (2021) Ensembl 2021. *Nucleic Acids Research*. 49(1), pp. 884–891. doi:10.1093/nar/gkaa942.

Lin, C.F., Naj, A.C. and Wang, L.S. (2013) Analyzing copy number variation using SNP array data: protocols for calling CNV and association tests. *Current Protocols in Human Genetics*. 79, pp. 1–17. doi:10.1002/0471142905.hg0127s79.

Macé, A., Tuke, M.A., Beckmann, J.S., *et al.* (2016) New quality measure for SNP array based CNV detection. *Bioinformatics*. 32(21), pp. 3298–3305. doi:10.1093/bioinformatics/btw477.

Nyangiri, O.A., Noyes, H., Mulindwa, J., *et al.* (2020) Copy number variation in human genomes from three major ethno-linguistic groups in Africa. *BMC Genomics*. 21(1), pp. 289. doi:10.1186/s12864-020-6669-y.

Png, G., Suveges, D., Park, Y., *et al.* (2019) Population-wide copy number variation calling using variant call format files from 6,898 individuals. *Genetic Epidemiology*. 44(1), pp. 79–89. doi:10.1002/gepi.22260.

Quick, C., Anugu, P., Musani, S., *et al.* (2020) Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. *Genetic Epidemiology*. 44(6), pp. 537–549. doi:10.1002/gepi.22326.

Ramsay, M., Crowther, N., Tambo, E., *et al.* (2016) H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Global Health, Epidemiology and Genomics*. 1(20), pp. 1–13. doi:10.1017/gheg.2016.17.

Romdhane, L., Mezzi, N., Dallali, H., *et al.* (2021) A map of copy number variations in the Tunisian population: a valuable tool for medical genomics in North Africa. *NPJ Genomic Medicine*. 6(1), pp. 3. doi:10.1038/s41525-020-00166-5.

Sham, P.C. and Purcell, S.M. (2014) Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*. 15(5), pp. 335–346. doi:10.1038/nrg3706.

Wang, K., Li, M., Hadley, D., *et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research*. 17(11), pp. 1665–1674. doi:10.1101/gr.6861907.

Yao, R., Yu, T., Qing, Y., *et al*. (2019) Evaluation of copy number detection from panel-based next-generation sequencing data. *Molecular Genetics and Genomic Medicine.* 7(1), pp. 1–8. doi: 10.1002/mgg3.513.

Zhang, Z., Cheng, H., Hong, X., *et al.* (2019) EnsembleCNV: an ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data. *Nucleic Acids Research*. 47(7), pp. 39. doi:10.1093/nar/gkz068.

Zhou, B., Ho, S.S., Zhang, X., *et al.* (2017) Whole-genome sequencing analysis of genomic copy number variation (CNV) using low-coverage and paired-end strategies is highly efficient and outperforms array based CNV analysis. *Journal of Medical Genetics*. 55(11), pp. 735–743. doi: 10.1136/jmedgenet-2018-105272

Zhou, Z., Wang, W., Wang, L.S., *et al.* (2018) Integrative DNA copy number detection and genotyping from sequencing and array-based platforms. *Bioinformatics*. 34(14), pp. 2349–2355. doi:10.1093/bioinformatics/bty104.

CHAPTER 5: OVERALL DISCUSSION AND CONCLUSION

5. CHAPTER 5: OVERALL DISCUSSION AND CONCLUSION

Various types of genetic variation exist in the human genome, ranging from small single nucleotide changes to large chromosomal changes (Eichler, 2019). In recent years, various studies have identified and revealed the importance of a class of sub-microscopic alterations known as copy number variation (CNV) (Freeman *et al.*, 2006; Redon *et al.*, 2006). While greatly understudied compared to other classes of variation, such as single nucleotide variation (SNV), CNV is now understood to contribute to a large proportion of human genetic diversity (Zhou *et al.*, 2018). In particular, genetic variation within genes involved in the absorption, distribution, metabolism and excretion (ADME) of therapeutic drugs has been shown to play a major role in how an individual responds to a particular drug. While SNVs in ADME genes have been extensively studied, a comprehensive analysis of CNVs affecting these genes is lacking, particularly in African populations, where the disease burden is high and adverse drug reactions (ADRs) are prevalent.

Various types of data can be used to call CNVs, including short-read sequencing data, long-read sequencing data and SNP array data, with each approach having its own limitations and advantages. The difficulty in calling structural variation (SV), including CNVs, is the main factor behind why this type of genetic variation remains understudied (De Coster and Van Broeckhoven, 2019). Additionally, generating population-wide CNV call sets remains challenging for various reasons. First of all, detecting CNVs from sequencing data requires a high depth of sequencing, which can make large, population-wide studies prohibitively expensive and thus restrict these types of studies to researchers with considerable funding and resources (Png *et al.*, 2019). This problem is exacerbated when it comes to long-read sequencing, which despite having been shown to result in more accurate CNV calling, is significantly more expensive than high-coverage, short-read WGS (Png *et al.*, 2019). Secondly, CNV calling poses a computation challenge, where processing hundreds of high-depth whole genome sequences (WGS) is demanding in terms of run time and memory requirements (Png *et al.*, 2019). A solution to both the cost and computational burden is calling CNVs from SNP array data that has already been generated for genome wide association studies (GWAS). However, WGS approaches are much more sensitive in terms of CNV calling when compared with the best-performing arrays (Zhou *et al.*, 2018). Lastly, over 100 different algorithms have been developed to call CNVs from

WGS data, each with their own weaknesses and advantages over other tools (Mahmoud *et al*., 2019). However, the results obtained from different tools vary greatly (Gabrielaite *et al.*, 2021) and the choice of tool is likely to have a great impact on the number and accuracy of CNV calls (Pabinger *et al.*, 2014; Tattini, D'Aurizio and Magi, 2015).

In this study we aimed to generate a population-wide CNV call set within ADME genes. This was achieved by calling CNVs from just under 1 000 high-coverage, short-read WGSs from individuals across sub-Saharan Africa (SSA), using a variety of different CNV calling tools (GenomeSTRiP, Manta, Delly, Lumpy and GATK) (Objective 1). The results from this work highlight the difficulties associated with CNV calling, as the tools used produced very different outputs, with only ~10% of CNVs being common across all tools. However, although the CNVs called by different tools were highly variable, the advantages of each tool were leveraged in order to obtain a more comprehensive dataset, and the merged dataset of 362 CNVs generated in this study provided important insights into the CNV landscape in ADME genes in SSA populations, allowing us to produce a list of ADME CNVs with potential relevance in an African setting (Objective 5).

Due to the limitations and difficulties of calling CNVs from short-read data, and the improvement in CNV calling accuracy that has been observed with the use of a long-read sequencing approach, we aimed to validate the CNVs called from short-read sequence data by generating long-read WGS data for a subset of eight individuals (Objective 2). To our knowledge, this is one of the first long-read WGS datasets to be generated from African individuals. When the short- and long-read CNVs were compared for the same eight individuals, approximately half of the short-read CNVs were concordant with the long-read data. Additionally, while only eight long-read sequences were generated, the results from this work illustrate that long-read sequencing is able to detect far more CNVs from far fewer samples compared to short-read approaches, as well as detecting a higher percentage of novel CNVs, thus providing invaluable insights into the CNV landscape in African populations. Lastly, this work shows that short-read sequencing alone is not adequate for accurate and comprehensive CNV calling and suggests that CNV calling in the future should make use of a combination of short- and long-read sequencing approaches (Objective 3).

As discussed above, the cost of WGS, both short- and long-read, may make this approach for population-wide CNV calling unfeasible in a resource-limited setting. An alternative approach is to exploit the data that has already been generated in GWAS studies, which can also be used for CNV calling. In this study, we called CNVs from the SNP array data of over 10 000 SSA individuals who had previously been genotyped on the Infinium H3Africa Consortium Array v2, in order to determine which CNVs could be called from the array, thus allowing for the assessment of its utility (Objective 4). The concordance between the array CNVs and the CNVs called from the short-read WGS data for the same 100 individuals was very low, indicating a limited utility to accurately call CNVs. However, when the array CNVs were compared to the CNVs called from the long-read WGS data for the same eight individuals, the concordance increased to over 50%, thus improving our confidence in the array results. However, very few CNVs were called from the array data compared to the short- and long-read WGS data, a known limitation of this approach, and this low resolution limits the utility of the array to generate a comprehensive population-wide CNV call set.

The CNVs identified in this study may be useful for precision medicine in the future. It is likely that we would be dependent on diagnostic tests that make use of a targeted approach to identify known CNVs that are associated with ADRs. Population data, such as the call set generated in this study, will be useful in designing these diagnostic tests. However, the high level of CNV diversity, as observed between different individuals in this study, may make designing a diagnostic test that is relevant to all populations and individuals difficult. In the case of unknown or novel CNVs, the results of this study have shown that, in a clinical setting, short-read sequencing would not be a viable approach for patient-specific testing. The fact that short-read sequencing alone is not adequate for accurate and comprehensive CNV calling would render it inappropriate as a diagnostic test to identify a mutation that may cause or be causing an ADR in a patient. High-coverage, long-read sequencing would be ideal in this case; however, the current high costs of long-read sequencing makes this an unfeasible approach in a clinical setting, particularly in resource-limited settings.

The following sections will discuss the limitations of this study, future work that could be done with the available data as well as extensions of this work with new data. Firstly, all of the CNVs identified in this study were called using *in silico* bioinformatics

approaches. Therefore, further laboratory work, such as droplet digital PCR (ddPCR), may be required in order to validate the presence of these CNVs as well as verify their breakpoints. Additionally, in the absence of further laboratory work to better understand the functional impact of these CNVs, as well as phenotypic data on how individuals with these CNVs respond to a given drug, we cannot fully understand the pharmacogenetic consequence that a CNV may have.

Quality control issues involving alternate contig regions of the BAM files used in this study were identified. However, due the high computational costs and large amount of time required to regenerate the BAM files and recall CNVs, CNVs within alternate contig regions were simply removed from the analysis. A disadvantage of this solution was that several important ADME genes could not be assessed. One of these genes is *CYP2D6* which is classified as a very important pharmacogene by PharmGKB and plays a role in the metabolism of ~25% of the drugs used today. Both deletions and duplications of *CYP2D6* have been previously reported and it would be interesting to assess the frequencies of these CNVs in a diverse set of SSA populations. The BAM files have since been regenerated and CNV recalling has begun.

Due to the very high cost of long-read sequencing and availability of DNA, we were only able to generate eight WGSs, all of which were obtained from individuals from South Africa. Therefore, future sequencing of additional samples from individuals across SSA will help to gain a better understanding of the CNV landscape in this region, across various populations and ethnolinguistic groups. Additionally, only one tool was used to call CNVs from the long-read sequencing data. Several other tools for calling CNVs from this type of data have been developed, including PBHoney (English, Salerno and Reid, 2014), SMRT-SV (Huddleston *et al.*, 2017) and Sniffles (Sedlazeck *et al.*, 2018). As each tool has its own limitations and advantages, running these additional tools on the data would allow for a comparison of the results obtained from the different tools as well as provide a more comprehensive and accurate CNV call set.

In this study the long reads generated by High Fidelity (HiFi) sequencing were aligned to a reference genome in order to generate the BAM files used for CNV calling. This means that the alignments produced and CNVs called are biased by the reference genome. As a solution to this, long-reads can be used to generate highly contiguous

de novo genome assemblies whereby reads are joined based on overlap in order to produce longer contiguous sequences without the need for a reference genome (Chin *et al.*, 2016). The use of this approach may potentially allow for more accurate CNV detection as well as the detection of larger CNVs from the long-read sequencing data.

Additionally, a method has recently been developed that combines long-read WGS with Strand-seq data in order to produce fully phased diploid genome assemblies (Ebert *et al.*, 2021). While these genome assemblies have been shown to improve SV calling, they can be used to interpret and call SVs in short-read datasets (Ebler *et al.*, 2020). Pangenome-based Genome Inference (PanGenIe) makes use of a panel of fully-phased assemblies to genotype all variants, including SVs, in short-read sequencing data (Ebler *et al.*, 2020). In this way, only a few, very expensive genome assemblies are required in order to genotype CNVs in much larger short-read datasets (Ebler *et al.*, 2020). With the addition of Strand-seq data, fully phased diploid genome assemblies could be generated for the eight samples with long-read sequencing data that has been generated in this study. These assemblies could then be used to call CNVs from the short-read data, and perhaps improve the CNV calls from this type of data.

Lastly, the CNV landscape of ADME genes has been studied in isolation, without taking into consideration other types of genetic variation within the same gene that may have functional effects. For example, an individual may have a duplication and therefore one would predict a rapid metaboliser or ultrarapid metaboliser phenotype. However, if the gene also contains SNVs that render it non-functional, the phenotypic prediction based only on the presence of a duplication, would be incorrect. Therefore, SNVs and CNVs in ADME genes need to be assessed together, and phased genome assemblies would help greatly in this regard.

Overall, this study has made use of several different approaches (SNP array, short-read WGS and long-read WGS) in order to provide insights into an understudied, yet potentially impactful, class of genetic variation. The CNVs identified in this study will be made available in a suitable database so that this data may be useful for future studies. Additionally, CNVs identified within ADME genes could be used to design an African-specific ADME CNV array, which could allow for the CNV calling from thousands of African individuals, thus improving our knowledge of the CNV landscape

in ADME genes. These CNVs may provide insight into the variability in drug response that is observed and may be useful for precision medicine in the future, allowing for improved efficacy and reduced ADRs, thus decreasing the burden of disease in SSA. The varied results obtained from the different approaches highlight the difficulties associated with CNV calling**.** Nevertheless, we have been able to provide the first assessment of the CNV landscape in ADME genes in SSA, as well as assess different approaches to CNV calling.

6. REFERENCES

Abdullahi, S.T., Soyinka, J.O., Olagunju, A., et al. (2020) CYP2B6*6 Genotype Specific Differences in Artemether-Lumefantrine Disposition in Healthy Volunteers. *The Journal of Clinical Pharmacology*. 60(3), pp. 351–360. doi:https://doi.org/10.1002/jcph.1527.

Abel, H.J., Larson, D.E., Regier, A.A., et al. (2020) Mapping and characterization of structural variation in 17,795 human genomes. *Nature*. 583(7814), pp. 83–89. doi:10.1038/s41586-020-2371-0.

Abyzov, A., Urban, A.E., Snyder, M., et al. (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*. 21(6), pp. 974–984. doi:10.1101/gr.114876.110.

Achan, J., Talisuna, A.O., Erhart, A., et al. (2011) Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malaria Journal*. 10, pp. 144. doi:10.1186/1475-2875-10-144.

Adedeji, A., Sanusi, B., Tella, A., et al. (2011) Exposure to anti-malarial drugs and monitoring of adverse drug reactions using toll-free mobile phone calls in private retail sector in Sagamu, Nigeria: Implications for pharmacovigilance. *Malaria Journal.* 10, pp. 230. doi:10.1186/1475-2875-10-230.

Ahmed, S., Zhou, Z., Zhou, J., et al. (2016) Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine. *Genomics, Proteomics and Bioinformatics.* 14(5), pp. 298–313. doi:10.1016/j.gpb.2016.03.008.

Aka, I., Bernal, C.J., Carroll, R., et al. (2017) Clinical Pharmacogenetics of Cytochrome P450-Associated Drugs in Children. *Journal of Personalized Medicine.* 7(4), pp. 14. doi:10.3390/jpm7040014.

Akhondzadeh, S. (2014) Personalized medicine: a tailor made medicine. *Avicenna Journal of Medical Biotechnology.* 6(4), pp. 191.

Al-Eitan, L.N., Almasri, A.Y. and Khasawneh, R.H. (2018) Impact of CYP2C9 and VKORC1 Polymorphisms on Warfarin Sensitivity and Responsiveness in Jordanian Cardiovascular Patients during the Initiation Therapy. *Genes*. 9(12), pp. 578. doi:10.3390/genes9120578.

Alessandrini, M. and Pepper, M.S. (2014) Priority pharmacogenetics for the African continent: focus on CYP450. *Pharmacogenomics.* 15(3), pp. 385–400. doi:10.2217/pgs.13.252.

Ali, S., Najmi, M.H., Tarning, J., et al. (2010) Pharmacokinetics of artemether and dihydroartemisinin in healthy Pakistani male volunteers treated with artemether-lumefantrine. *Malaria Journal*. 9, pp. 275. doi:10.1186/1475-2875-9-275.

Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping. *Nature Reviews Genetics*. 12(5), pp. 363–376. doi:10.1038/nrg2958.

Allied Against Cancer (2021) Cancer Access Partnership. Available at: https://www.alliedagainstcancer.org/access-partnership (Accessed: 2 May 2021).

Allocati, N., Masulli, M., Di Ilio, C., et al. (2018) Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases. *Oncogenesis*. 7(1), pp. 8. doi:10.1038/s41389-017-0025-3.

Amarasinghe, S.L., Su, S., Dong, X., et al. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*. 21(1), pp. 30. doi:10.1186/s13059-020-1935-5.

Appiah, B. (2012) Africa struggles to improve drug safety. *Canadian Medical Association journal.* 184(10), pp. 533–534. doi:10.1503/cmaj.109-4199.

Arbex, M.A., Varella, M. C.L., de Siqueira, H.R., et al. (2010) Antituberculosis drugs: drug interactions, adverse effects, and use in special situations. Part 1: first-line drugs. *Jornal Brasileiro de Pneumologia*. 36, pp. 626–640. doi:10.1590/s1806-37132010000500016.

Ardui, S., Ameur, A., Vermeesch, J.R., et al. (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*. 46(5), pp. 2159–2168. doi:10.1093/nar/gky066.

Arlt, M.F., Wilson, T.E. and Glover, T.W. (2012) Replication stress and mechanisms of CNV formation. *Current Opinion in Genetics & Development*. 22(3), pp. 204–210. doi:10.1016/j.gde.2012.01.009.

Avert (2020) *HIV and AIDS in East and Southern Africa Regional Overview*. Available at: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/overview (Accessed: 3 November 2019).

Bahnassy, A.A., Abdellateif, M.S. and Zekri, A.N. (2020) Cancer in Africa: Is It a Genetic or Environmental Health Problem? *Frontiers in Oncology*. 10, pp. 2508. doi:10.3389/fonc.2020.604214.

Bains, R.K. (2013) African variation at Cytochrome P450 genes: Evolutionary aspects and the implications for the treatment of infectious diseases. *Evolution, Medicine, and Public Health.* 2013(1), pp. 118–134. doi:10.1093/emph/eot010.

Bibi, Z. (2008) Role of cytochrome P450 in drug interactions. *Nutrition & Metabolism*. 5(1), pp. 27. doi:10.1186/1743-7075-5-27.

Botton, M.R., Lu, X., Zhao, G., et al. (2019) Structural variation at the CYP2C locus: Characterization of deletion and duplication alleles. *Human Mutation*. 40(11), pp. 37–51. doi:10.1002/humu.23855.

Boyle, P., Ngoma, T., Sullivan, R., et al. (2019) Cancer in Africa: the way forward. *Ecancer Medical Science*. 13, pp. 953. doi:10.3332/ecancer.2019.953.

Bradford, L.D. (2002) CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants. *Pharmacogenomics.* 3(2), pp. 229–243. doi:10.1517/14622416.3.2.229.

Brito, T.C., Possuelo, L.G., Valim, A.R.M., et al. (2014) Polymorphisms in CYP2E1, GSTM1 and GSTT1 and anti-tuberculosis drug-induced hepatotoxicity. *Anais da Academia Brasileira de Ciências.* 86, pp. 855–865. doi: 10.1590/0001-3765201420130350.

Buysse, K., Delle Chiaie, B., Van Coster, R., et al. (2009) Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *European Journal of Medical Genetics*. 52(6), pp. 398–403. doi:10.1016/j.ejmg.2009.09.002.

Cabianca, D.S. and Gabellini, D. (2010) FSHD: copy number variations on the theme of muscular dystrophy. *The Journal of Cell Biology*. 191(6), pp. 1049–1060. doi:10.1083/jcb.201007028.

Campbell, M.C. and Tishkoff, S.A. (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*. 9, pp. 403–433. doi:10.1146/annurev.genom.9.081307.164258.

Carter, N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*. 39(7), pp. 16–21. doi:10.1038/ng2028.

Castellino, S., Moss, L., Wagner, D., et al. (2013) Metabolism, excretion, and mass balance of the HIV-1 integrase inhibitor dolutegravir in humans. *Antimicrobial Agents and Chemotherapy.* 57(8), pp. 3536–3546. doi:10.1128/AAC.00292-13.

Caudle, K.E., Dunnenberger, H.M., Freimuth, R.R., et al. (2017) Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genetics in Medicine*. 19(2), pp. 215–223. doi:10.1038/gim.2016.87.

Chaisson, M.J.P., Wilson, R.K. and Eichler, E.E. (2015) Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*. 16(11), pp. 627–640. doi:10.1038/nrg3933.

Chen, K., Wallis, J.W., McLellan, M.D., et al. (2009) BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*. 6(9), pp. 677–681. doi:10.1038/nmeth.1363.

Chen, X., Schulz-Trieglaff, O., Shaw, R., et al. (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 32(8), pp. 1220–1222. doi:10.1093/bioinformatics/btv710.

Chin, C.S., Peluso, P., Sedlazeck, F.J., et al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), pp. 1050–1054. doi:10.1038/nmeth.4035.

Chirehwa, M.T., McIlleron, H., Rustomjee, R., et al. (2017) Pharmacokinetics of Pyrazinamide and Optimal Dosing Regimens for Drug-Sensitive and -Resistant Tuberculosis. *Antimicrobial Agents and Chemotherapy*. 61(8), pp. 2171–2179. doi:10.1128/AAC.00490-17.

Cho, D.Y., Shen, J.H.Q., Lemler, S.M., et al. (2016) Rifampin enhances cytochrome P450 (CYP) 2B6-mediated efavirenz 8-hydroxylation in healthy volunteers. *Drug Metabolism and Pharmacokinetics.* 31(2), pp. 107–116. doi:10.1016/j.dmpk.2015.07.002.

Choudhury, A., Aron, S., Sengupta, D., et al. (2018) African genetic diversity provides novel insights into evolutionary history and local adaptations. *Human Molecular Genetics*. 27, pp. 209–218. doi:10.1093/hmg/ddy161.

Ciccacci, C., Latini, A., Politi, C., et al. (2017) Impact of glutathione transferases genes polymorphisms in nevirapine adverse reactions: a possible role for GSTM1 in SJS/TEN susceptibility. *European Journal of Clinical Pharmacology*. 73(10), pp. 1253–1259. doi:10.1007/s00228-017-2295-2.

Clarke, J., Wu, H.C., Jayasinghe, L., et al. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*. 4(4), pp. 265–270. doi:10.1038/nnano.2009.12.

Clinton Health Access Initiative (2020) *New agreements to lower prices and increase access to lifesaving cancer treatment in sub-Saharan Africa*. Available at: https://www.clintonhealthaccess.org/agreements-to-increase-access-to-cancer-treatment-in-africa/ (Accessed: 2 May 2021).

Colella, S., Yau, C., Taylor, J.M., et al. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*. 35(6), pp. 2013–2025. doi:10.1093/nar/gkm076.

Collins, R.L., Brand, H., Karczewski, K.J., et al. (2020) A structural variation reference for medical and population genetics. *Nature*. 581(7809), pp. 444–451. doi:10.1038/s41586-020-2287-8.

Conrad, D.F., Andrews, T.D., Carter, N.P., et al. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics*. 38(1), pp. 75–81. doi:10.1038/ng1697.

Conrad, M.D. and Rosenthal, P.J. (2019) Antimalarial drug resistance in Africa: the calm before the storm? *The Lancet Infectious Diseases*. 19(10), pp. 338–351. doi:10.1016/S1473-3099(19)30261-0.

de Coster, W. and Van Broeckhoven, C. (2019) Newest Methods for Detecting Structural Variations. *Trends in Biotechnology*. 37(9), pp. 973–982. doi:https://doi.org/10.1016/j.tibtech.2019.02.003.

Coutelier, M., Holtgrewe, M., Jäger, M., et al. (2021) Combining callers improves the detection of copy number variants from whole-genome sequencing. *European Journal of Human Genetics.* 30, pp. 178–186. doi:10.1038/s41431-021-00983-x.

Crews, K.R., Hicks, J.K., Pui, C.H., et al. (2012) Pharmacogenomics and individualized medicine: translating science into practice. *Clinical Pharmacology and Therapeutics.* 92(4), pp. 467–475. doi:10.1038/clpt.2012.120.

Cronin-Fenton, D.P., Damkier, P. and Lash, T.L. (2014) Metabolism and transport of tamoxifen in relation to its effectiveness: new perspectives on an ongoing controversy. *Future Oncology*. 10(1), pp. 107–122. doi:10.2217/fon.13.168.

Dandara, C., Swart, M., Mpeta, B., et al. (2014) Cytochrome P450 pharmacogenetics in African populations: implications for public health. *Expert Opinion on Drug Metabolism & Toxicology.* 10(6), pp. 769–785. doi:10.1517/17425255.2014.894020.

Dean, L. and Kane, M. (2016) 'Capecitabine Therapy and DPYD Genotype' in Pratt, V.M., Scott, S.A., and Pirmohamed, M., et al. (eds.) *Medical Genetics Summaries* [Internet]. Bethesda: National Center for Biotechnology Information.

DecodeGenetics (2019) *svimmer*. Available at: https://github.com/DecodeGenetics/svimmer (Accessed: 22 June 2021).

Delaneau, O., Howie, B., Cox, A.J., et al. (2013) Haplotype estimation using sequencing reads. *American Journal of Human Genetics*. 93(4), pp. 687–696. doi:10.1016/j.ajhg.2013.09.002.

Dent, J., Manner, C.K., Milner, D., et al. (2017) *Africa's Emerging Cancer Crisis: A Call to Action*. Available at: https://bvgh.org/wp-content/uploads/2017/07/Africas-Emerging-Cancer-Crisis-A-Call-to-Action.pdf (Accessed: 2 May 2021).

Dhoro, M., Zvada, S., Ngara, B., et al. (2015) CYP2B6*6, CYP2B6*18, Body weight and sex are predictors of efavirenz pharmacokinetics and treatment response: population pharmacokinetic modeling in an HIV/AIDS and TB cohort in Zimbabwe. *BMC Pharmacology & Toxicology*. 16, pp. 4. doi:10.1186/s40360-015-0004-2.

Dodgen, T.M., Hochfeld, W.E., Fickl, H., et al. (2013) Introduction of the AmpliChip CYP450 Test to a South African cohort: A platform comparative prospective cohort study. *BMC Medical Genetics*. 14(1). pp. 1–15. doi:10.1186/1471-2350-14-20.

Dyer, O. (2017) Pfizer and Cipla plan to make cancer drugs affordable in Africa. *BMJ*. 359, pp. 1, doi:10.1136/bmj.j4848.

Ebert, P., Audano, P.A., Zhu, Q., et al. (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 372(6537), pp. 1–33. doi:10.1126/science.abf7117.

Ebler, J., Clarke, W.E., Rausch, T., et al. (2020) Pangenome-based genome inference. *bioRxiv*. doi:10.1101/2020.11.11.378133.

Eggertsson, H.P., Kristmundsdottir, S., Beyter, D., et al. (2019) GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*. 10(1), pp. 5402. doi:10.1038/s41467-019-13341-9.

Eichler, E.E. (2019) Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *The New England Journal of Medicine*. 381(1), pp. 64–74. doi:10.1056/NEJMra1809315.

Eid, J., Fehr, A., Gray, J., et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science*. 323(5910), pp. 133–138. doi:10.1126/science.1162986.

Elliot, E., Mahungu, T. and Owen, A. (2017) 'Current Progress in the Pharmacogenetics of Infectious Disease Therapy' in Tibayrenc, M. (ed.) *Genetics and Evolution of Infectious Disease*. London: Elsevier. pp. 435–457. doi:https://doi.org/10.1016/B978-0-12-799942-5.00019-6.

English, A.C., Salerno, W.J. and Reid, J.G. (2014) PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*. 15, pp. 180. doi:10.1186/1471-2105-15-180.

Feuk, L., Carson, A.R. and Scherer, S.W. (2006) Structural variation in the human genome. *Nature Reviews Genetics*. 7(2), pp. 85–97. doi:10.1038/nrg1767.

Fiegler, H., Redon, R., Andrews, D., et al. (2006) Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Research*. 2006/11/22. 16(12), pp. 1566–1574. doi:10.1101/gr.5630906.

Flegg, J.A., Metcalf, C.J.E., Gharbi, M., et al. (2013) Trends in antimalarial drug use in Africa. *The American Journal of Tropical Medicine and Hygiene*. 89(5), pp. 857–865. doi:10.4269/ajtmh.13-0129.

Freeman, J.L., Perry, G.H., Feuk, L., et al. (2006) Copy number variation: New insights in genome diversity. *Genome Research*. 16(8), pp. 949–961. doi:10.1101/gr.3677206.

Fuselli, S., de Filippo, C., Mona, S., et al. (2010) Evolution of detoxifying systems: the role of environment and population history in shaping genetic diversity at human CYP2D6 locus. *Pharmacogenetics and Genomics*. 20(8), pp. 485–499. doi:10.1097/FPC.0b013e32833bba25.

Gabrielaite, M., Torp, M.H., Rasmussen, M.S., et al. (2021) A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers*. 13(24), pp. 1–21. doi:2021.04.30.442110. doi:10.1101/2021.04.30.442110.

Gaedigk, A. and Leeder, J.S. (2012) CYP2D6, SULT1A1 and UGT2B17 copy number variation: quantitative detection by multiplex PCR. *Pharmacogenomics*. 13, pp. 91–111. doi: 10.2217/pgs.11.135.

Gallagher, C.J., Kadlubar, F.F., Muscat, J.E., et al. (2007) The UGT2B17 gene deletion polymorphism and risk of prostate cancer. A case-control study in Caucasians. *Cancer Detection and Prevention*. 31(4), pp. 310–315. doi:10.1016/j.cdp.2007.07.005.

Garte, S., Gaspari, L., Alexandrie, A.K., et al. (2001) Metabolic gene polymorphism frequencies in control populations. *Cancer Epidemiology, Biomarkers & Prevention*. 10(12), pp. 1239–1248.

Gebreyohannes, E.A., Bhagavathula, A.S., Seid, M.A., et al. (2017) Antimalarial Treatment Outcomes In Ethiopia: A Systematic Review And Meta-Analysis. *Value in Health*. 20(9), pp. 795. doi:10.1016/j.jval.2017.08.2346.

Genovese, G., Handsaker, R.E., Li, H., et al. (2013) Using population admixture to help complete maps of the human genome. *Nature Genetics*. 45(4), pp. 406–414. doi:10.1038/ng.2565.

Glusman, G., Severson, A., Dhankani, V., et al. (2015) Identification of copy number variants in whole-genome data using Reference Coverage Profiles. *Frontiers in Genetics*. 6(45), pp. 1–13. doi: 10.3389/fgene.2015.00045.

Goh, L.L., Lim, C.W., Sim, W.C., et al. (2017) Analysis of Genetic Variation in CYP450 Genes for Clinical Implementation. *PLoS One*. 12(1), pp. 1–16. doi:10.1371/journal.pone.0169233.

Gonzaga-Jauregui, C., Lupski, J. and Gibbs, R. (2012) Human Genome Sequencing in Health and Disease. *Annual Review of Medicine*. 63, pp. 35–61. doi:10.1146/annurev-med-051010-162644.

Griese, E.U., Asante-Poku, S., Ofori-Adjei, D., et al. (1999) Analysis of the CYP2D6 gene mutations and their consequences for enzyme function in a West African population. *Pharmacogenetics*. 9(6), pp. 715–723.

Grimwood, J., Gordon, L.A., Olsen, A., et al. (2004) The DNA sequence and biology of human chromosome 19. *Nature*. 428(6982), pp. 529–535. doi:10.1038/nature02399.

Gu, W., Zhang, F. and Lupski, J. (2008) Mechanisms for human genomic rearrangements. *PathoGenetics*. 1, pp. 4. doi:10.1186/1755-8417-1-4.

Habibi, M., Mirfakhraie, R., Khani, M., et al. (2017) Genetic variations in UGT2B28, UGT2B17, UGT2B15 genes and the risk of prostate cancer: A case-control study. *Gene*. 634, pp. 47–52. doi:https://doi.org/10.1016/j.gene.2017.08.038.

Handsaker, R.E., Van Doren, V., Berman, J.R., et al. (2015) Large multiallelic copy number variations in humans. *Nature Genetics*. 47(3), pp. 296–303. doi:10.1038/ng.3200.

Hastings, P.J., Ira, G. and Lupski, J.R. (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics*. 2009/01/30. 5(1), pp. 1–9. doi:10.1371/journal.pgen.1000327.

He, X. and Feng, S. (2015) Role of Metabolic Enzymes P450 (CYP) on Activating Procarcinogen and their Polymorphisms on the Risk of Cancers. *Current Drug Metabolism*. 16(10), pp. 850–863. doi:10.2174/1389200216101512101645O1.

He, Y., Hoskins, J.M. and McLeod, H.L. (2011) Copy number variants in pharmacogenetic genes. *Trends in Molecular Medicine*. 17(5), pp. 244–251. doi:10.1016/j.molmed.2011.01.007.

Hebbring, S.J., Adjei, A.A., Baer, J.L., et al. (2007) Human SULT1A1 gene: copy number differences and functional implications. *Human Molecular Genetics*. 16(5), pp. 463–470. doi:10.1093/hmg/ddl468.

HIV.gov (2021) *Glossary of HIV/AIDS-Related Terms*. Available at: https://clinicalinfo.hiv.gov/sites/default/files/glossary/Glossary-English_HIVinfo.pdf (Accessed: 30 April 2021).

Hon, T., Mars, K., Young, G., et al. (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*. 7(1), pp. 399. doi:10.1038/s41597-020-00743-4.

Hovelson, D.H., Xue, Z., Zawistowski, M., et al. (2017) Characterization of ADME gene variation in 21 populations by exome sequencing. *Pharmacogenetics and Genomics*. 27(3), pp. 89–100. doi:10.1097/FPC.0000000000000260.

Howe, K.L., Achuthan, P., Allen, James, et al. (2021) Ensembl 2021. *Nucleic Acids Research*. 49(1), pp. 884–891. doi:10.1093/nar/gkaa942.

Huddleston, J., Chaisson, M.J.P., Steinberg, K.M., et al. (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*. 27(5), pp. 677–685. doi:10.1101/gr.214007.116.

Ilett, K.F., Ethell, B.T., Maggs, J.L., et al. (2002) Glucuronidation of dihydroartemisinin in vivo and by human liver microsomes and expressed UDP-glucuronosyltransferases. *Drug Metabolism and Disposition*. 30(9), pp. 1005–1012. doi:10.1124/dmd.30.9.1005.

Ingelman-Sundberg, M. (2004) Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. *Trends in Pharmacological Sciences*. 25(4), pp. 193–200. doi:https://doi.org/10.1016/j.tips.2004.02.007.

Ingelman-Sundberg, M. (2005) Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *The Pharmacogenomics Journal*. 5(1), pp. 6–13. doi:10.1038/sj.tpj.6500285.

Ingelman-Sundberg, M., Mkrtchian, S., Zhou, Y., et al. (2018) Integrating rare genetic variants into pharmacogenetic drug response predictions. *Human Genomics*. 12(1), p. 26. doi:10.1186/s40246-018-0157-3.

Ionita-laza, I., Rogers, A.J., Lange, C., et al. (2010) Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*. 93(1), pp. 22–26. doi:10.1016/j.ygeno.2008.08.012.

Ishiguro, A., Kubota, T., Ishikawa, H., et al. (2004) Metabolic activity of dextromethorphan O-demethylation in healthy Japanese volunteers carrying duplicated CYP2D6 genes: duplicated allele of CYP2D6*10 does not increase CYP2D6 metabolic activity. *International Journal of Clinical Chemistry*. 344(1–2), pp. 201–204. doi:10.1016/j.cccn.2004.03.002.

Ito, T., Tsuji, M., Mori, Y., et al. (2015) Effect of CYP2A6*4 genetic polymorphisms on smoking behaviours and nicotine dependence in a general population of Japanese men. *Fukushima Journal of Medical Science*. 61(2), pp. 125–130. doi:10.5387/fms.2015-14.

Jakubosky, D., Smith, E.N., D'Antonio, M., et al. (2020) Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. *Nature Communications*. 11(1), pp. 2928. doi:10.1038/s41467-020-16481-5.

Jarvis, J.P., Peter, A.P. and Shaman, J.A. (2019) Consequences of CYP2D6 Copy-Number Variation for *Pharmacogenomics in Psychiatry*. Frontiers in psychiatry. 10, pp. 432. doi:10.3389/fpsyt.2019.00432.

Jiang, Y., Wang, Y. and Brudno, M. (2012) PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*. 28(20), pp. 2576–2583. doi:10.1093/bioinformatics/bts484.

Johansson, I. and Ingelman-Sundberg, M. (2009) CNVs of human genes and their implication in pharmacogenetics. *Cytogenetic and Genome Research*. 123(1–4), pp. 195–204. doi:10.1159/000184709.

Kalow, W., Tang, B.K. and Endrenyi, L. (1998) Hypothesis: comparisons of inter- and intra-individual variations can substitute for twin studies in drug research. *Pharmacogenetics*. 8(4), pp. 283–289. doi:10.1097/00008571-199808000-00001.

Kamataki, T., Fujieda, M., Kiyotani, K., et al. (2005) Genetic polymorphism of CYP2A6 as one of the potential determinants of tobacco-related cancer risk. *Biochemical and Biophysical Research Communications*. 338(1), pp. 306–310. doi:10.1016/j.bbrc.2005.08.268.

Kassahun, K., McIntosh, I., Cui, D., et al. (2007) Metabolism and disposition in humans of raltegravir (MK-0518), an anti-AIDS drug targeting the human immunodeficiency virus 1 integrase enzyme. *Drug Metabolism and Disposition*. 35(9), pp. 1657–1663. doi:10.1124/dmd.107.016196.

Kenmotsu, H. and Tanigawara, Y. (2015) Pharmacokinetics, dynamics and toxicity of docetaxel: Why the Japanese dose differs from the Western dose. *Cancer Science*. 106(5), pp. 497–504. doi:10.1111/cas.12647.

Kis, O., Sankaran-Walters, S., Hoque, M.T., et al. (2016) HIV-1 Alters Intestinal Expression of Drug Transporters and Metabolic Enzymes: Implications for Antiretroviral Drug Disposition. *Antimicrobial Agents and Chemotherapy*. 60(5), pp. 2771–2781. doi:10.1128/AAC.02278-15.

Klambauer, G., Schwarzbauer, K., Mayr, A., et al. (2012) cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research*. 40(9), pp. 1–14. doi:10.1093/nar/gks003.

Koboldt, D.C., Larson, D.E., Chen, K., et al. (2012) Massively parallel sequencing approaches for characterization of structural variation. *Methods in Molecular Biology*. 838, pp. 369–384. doi:10.1007/978-1-61779-507-7_18.

Korbel, J.O., Abyzov, A., Mu, X.J., et al. (2009) PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*. 10(2), pp. 1–14. doi:10.1186/gb-2009-10-2-r23.

Korhonova, M., Doricakova, A. and Dvorak, Z. (2015) Optical Isomers of Atorvastatin, Rosuvastatin and Fluvastatin Enantiospecifically Activate Pregnane X Receptor PXR and Induce CYP2A6, CYP2B6 and CYP3A4 in Human Hepatocytes. *PLoS One*. 10(9), pp. 1–18. doi:10.1371/journal.pone.0137720.

Kosugi, S., Momozawa, Y., Liu, X., et al. (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*. 20(1), pp. 117. doi:10.1186/s13059-019-1720-5.

Laäs, D.J. and Naidoo, M. (2018) An evaluation of warfarin use at an urban district-level hospital in KwaZulu-Natal Province, South Africa. *South African Medical Journal*. 108(12), pp. 1046–1050. doi: 10.7196/SAMJ.2018.v108i12.13256.

Langmia, I.M., Just, K.S., Yamoune, S., et al. (2021) CYP2B6 Functional Variability in Drug Metabolism and Exposure Across Populations – Implication for Drug Safety, Dosing, and Individualized Therapy. *Frontiers in Genetics*. 12(12), pp. 1-–21. doi: 10.3389/fgene.2021.692234.

Layer, R.M., Chiang, C., Quinlan, A.R., et al. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*. 15(6), pp. 84. doi:10.1186/gb-2014-15-6-r84.

Lee, J.H. and Jeon, J.T. (2008) Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels. *Cytogenetic and Genome Research*. 123(1–4), pp. 333–342. doi:10.1159/000184725.

Lentz, F., Tran, A., Rey, E., et al. (2005) Pharmacogenomics of fluorouracil, irinotecan, and oxaliplatin in hepatic metastases of colorectal cancer: Clinical implications. *American Journal of Pharmacogenomics*. 5(1), pp. 21–33. doi:10.2165/00129785-200505010-00002.

Li, Q. and Weina, P. (2010) Artesunate: The Best Drug in the Treatment of Severe and Complicated Malaria. *Pharmaceuticals*. 3(7), pp. 2322–2332. doi:10.3390/ph3072322.

Lim, L. and Burge, C. (2001) A computational analysis of sequence features involved in recognition of short introns. *PNAS.* 98(20), pp. 11193–11198.

Lin, C.F., Naj, A.C. and Wang, L.S. (2013) Analyzing copy number variation using SNP array data: protocols for calling CNV and association tests. *Current Protocols in Human Genetics*. 79, pp. 1–17. doi:10.1002/0471142905.hg0127s79.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., et al. (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*. 42, pp. 986–992. doi:10.1093/nar/gkt958.

Macé, A., Tuke, M.A., Beckmann, J.S., et al. (2016) New quality measure for SNP array based CNV detection. *Bioinformatics*. 32(21), pp. 3298–3305. doi:10.1093/bioinformatics/btw477.

Maheshwari, R., Kuche, K., Mane, A., et al. (2018) 'Manipulation of Physiological Processes for Pharmaceutical Product Development' in Tekade, R.K. (ed.) *Dosage Form Design Consideration*. Massachusetts: Academic Press, pp. 701–729.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., et al. (2019) Structural variant calling: the long and the short of it. *Genome Biology*. 20(1), pp. 246. doi:10.1186/s13059-019-1828-7.

Manosuthi, W., Wiboonchutikul, S. and Sungkanuparph, S. (2016) Integrated therapy for HIV and tuberculosis. *AIDS Research and Therapy*. 13(1), pp. 22. doi:10.1186/s12981-016-0106-y.

Mantere, T., Kersten, S. and Hoischen, A. (2019) Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in Genetics*. 10, pp. 426. doi:10.3389/fgene.2019.00426.

Marcath, L.A., Kidwell, K.M., Robinson, A.C., et al. (2019) Patients carrying CYP2C8*3 have shorter systemic paclitaxel exposure. *Pharmacogenomics*. 20(2), pp. 95–104. doi:10.2217/pgs-2018-0162.

Martin, A.R., Teferra, S., Möller, M., et al. (2018) The critical needs and challenges for genetic architecture studies in Africa. *Current Opinion in Genetics & Development*. 53, pp. 113–120. doi:10.1016/j.gde.2018.08.005.

Martis, S., Mei, H., Vijzelaar, R., et al. (2013) Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. *The Pharmacogenomics Journal*. 13(6), pp. 558–566. doi:10.1038/tpj.2012.48.

Matimba, A., Dhoro, M. and Dandara, C. (2016) Is there a role of pharmacogenomics in Africa. *Global Health, Epidemiology and Genomics*. 1, pp. 9. doi:10.1017/gheg.2016.4.

Mayosi, B.M., Flisher, A.J., Lalloo, U.G., et al. (2009) The burden of non-communicable diseases in South Africa. *Lancet*. 374(9693), pp. 934–947. doi:10.1016/S0140-6736(09)61087-4.

McCarroll, S.A., Hadnott, T.N., Perry, G.H., et al. (2006) Common deletion polymorphisms in the human genome. *Nature Genetics*. 38(1), pp. 86–92. doi:10.1038/ng1696.

Mehrotra, N., Gupta, M., Kovar, A., et al. (2007) The role of pharmacokinetics and pharmacodynamics in phosphodiesterase-5 inhibitor therapy. *International Journal of Impotence Research*. 19(3), pp. 253–264. doi:10.1038/sj.ijir.3901522.

Meintjes, G., Moorhouse, M.A., Carmona, S., et al. (2017) Adult antiretroviral therapy guidelines 2017. *Southern African Journal of HIV Medicine*. 18(1), pp. 776. doi:10.4102/sajhivmed.v18i1.776.

Ménard, V., Eap, O., Harvey, M., et al. (2009) Copy-number variations (CNVs) of the human sex steroid metabolizing genes UGT2B17 and UGT2B28 and their associations with a UGT2B15 functional polymorphism. *Human Mutation*. 30(9), pp. 1310–1319. doi:https://doi.org/10.1002/humu.21054.

Meyerson, M., Gabriel, S. and Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*. 11(10), pp. 685–696. doi:10.1038/nrg2841.

Mills, R.E., Walter, K., Stewart, C., et al. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*. 470, pp. 59–65. doi:10.1038/nature09708.Mapping.

Mitchell, C., Gregersen, N. and Krause, A. (2011) Novel CYP2C9 and VKORC1 gene variants associated with warfarin dosage variability in the South African black population. *Pharmacogenomics*. 12(7), pp. 953–63. doi:10.2217/pgs.11.36.

Mittal, B., Tulsyan, S., Kumar, S., et al. (2015) 'Cytochrome P450 in Cancer Susceptibility and Treatment' in Makowski, G. (ed.) *Advances in Clinical Chemistry*. Amsterdam: Elsevier. pp. 77–139. doi:https://doi.org/10.1016/bs.acc.2015.06.003.

Mroziewicz, M. and Tyndale, R. (2010) Pharmacogenetics: a tool for identifying genetic factors in drug dependence and response to treatment. *Addiction Science and Clinical Practice*. 5(2), pp. 17–29.

Mukonzo, J.K., Owen, J.S., Ogwal-Okeng, J., et al. (2014) Pharmacogenetic-based efavirenz dose modification: suggestions for an African population and the different CYP2B6 genotypes. *PLoS One*. 9(1), pp. 1–8. doi:10.1371/journal.pone.0086919.

Mulder, N., Lombard, Z., Owolabi, M.O., et al. (2020) The Genetic and Environmental Basis for Diseases in Understudied Populations. *Frontiers in Genetics*. 11(559956), pp. 1–2. doi: 10.3389/fgene.2020.559956.

Navaratnam, V., Mansor, S.M., Sit, N.W., et al. (2000) Pharmacokinetics of artemisinin-type compounds. *Clinical Pharmacokinetics*. 39(4), pp. 255–270. doi:10.2165/00003088-200039040-00002.

Ndagije, H.B., Nambasa, V., Manirakiza, L., et al. (2018) The Burden of Adverse Drug Reactions Due to Artemisinin-Based Antimalarial Treatment in Selected Ugandan Health Facilities: An Active Follow-Up Study. *Drug Safety*. 41(8), pp. 753–765. doi:10.1007/s40264-018-0659-x.

Nguyen, D.Q., Webber, C. and Ponting, C.P. (2006) Bias of Selection on Human Copy-Number Variants. *PLoS Genetics*. 2(2), pp. 197–207. doi: 10.1371/journal.pgen.0020020

Nijkamp, J.F., Van Den Broek, M.A., Geertman, J.M.A., et al. (2012) De novo detection of copy number variation by co-assembly. *Bioinformatics*. 28(24), pp. 3195–3202. doi:10.1093/bioinformatics/bts601.

Nowakowska, B. (2017) Clinical interpretation of copy number variants in the human genome. *Journal of Applied Genetics*. 58(4), pp. 449–457. doi:10.1007/s13353-017-0407-4.

Nyangiri, O.A., Noyes, H., Mulindwa, J., et al. (2020) Copy number variation in human genomes from three major ethno-linguistic groups in Africa. *BMC Genomics*. 21(1), pp. 289. doi:10.1186/s12864-020-6669-y.

Nyirenda, M.J. (2016) Non-communicable diseases in sub-Saharan Africa: Understanding the drivers of the epidemic to inform intervention strategies. *International Health*. 8(3), pp. 157–158. doi:10.1093/inthealth/ihw021.

Ogutu, B. (2013) Artemether and lumefantrine for the treatment of uncomplicated Plasmodium falciparum malaria in sub-Saharan Africa. *Expert Opinion on Pharmacotherapy*. 14(5), pp. 643–654. doi:10.1517/14656566.2013.771167.

Olsson, A.G., McTaggart, F. and Raza, A. (2002) Rosuvastatin: a highly effective new HMG-CoA reductase inhibitor. *Cardiovascular Drug Reviews*. 20(4), pp. 303–328. doi:10.1111/j.1527-3466.2002.tb00099.x.

Pabinger, S., Dander, A., Fischer, M., et al. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*. 15(2), pp. 256–278. doi:10.1093/bib/bbs086.

PacBio (2020) *Sequencing 101: The Evolution of DNA Sequencing Tools*. Available at: https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/ (Accessed: 01/10/2020).

Pacey, S., Workman, P. and Sarker, D. (2011) 'Pharmacokinetics and Pharmacodynamics in Drug Development' in Schwab, M. (ed.) *Encyclopedia of Cancer*. Berlin: Springer, pp. 2845–2848. doi:10.1007/978-3-642-16483-5_4501.

Pacific Biosciences (2017) pbsv. Available at: https://github.com/PacificBiosciences/pbsv (Accessed: 3 May 2021).

Paul, A. (2019) 'Drug Absorption and Bioavailability' in Raj, G.M. and Raveendran, R. (eds.) *Introduction to Basics of Pharmacology and Toxicology*. Singapore: Springer, pp. 81–88. doi:10.1007/978-981-32-9779-1_5.

PharmaADME.org (2021) *PharmaADME.org*. Available at: http://pharmaadme.org. (Accessed: 15 June 2019)

Phillips, A.N., Venter, F., Havlir, D., et al. (2019) Risks and benefits of dolutegravir-based antiretroviral drug regimens in sub-Saharan Africa: a modelling study. *The Lancet*. 6(2), pp. 116–127. doi:10.1016/S2352-3018(18)30317-5.

Piacentini, S., Polimanti, R., Porreca, F., et al. (2011) GSTT1 and GSTM1 gene polymorphisms in European and African populations. *Molecular Biology Reports*.

38(2), pp. 1225–1230. doi:10.1007/s11033-010-0221-0.
Pinto, D., Darvishi, K., Shi, X., et al. (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*. 29(6), pp. 512–520. doi:10.1038/nbt.1852.
Pirooznia, M., Goes, F. and Zandi, P.P. (2015) Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics*. 6(138), pp. 1–9. doi:10.3389/fgene.2015.00138.
Png, G., Suveges, D., Park, Y., et al. (2019) Population-wide copy number variation calling using variant call format files from 6,898 individuals. *Genetic Epidemiology*. 44(1), pp. 79–89. doi:10.1002/gepi.22260.
Poplin, R., Ruano-Rubio, V., DePristo, M.A., et al. (2018) Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. doi:10.1101/201178.
Pös, O., Radvanszky, J., Buglyó, G., *et al*. (2021) DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal*. 44(5), pp. 548–559. doi: 10.1016/j.bj.2021.02.003.
Quick, C., Anugu, P., Musani, S., et al. (2020) Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. *Genetic Epidemiology*. 44(6), pp. 537–549. doi:10.1002/gepi.22326.
Rajman, I., Knapp, L. and Hanna, I. (2020) Genetic Diversity in Drug Transporters: Impact in African Populations. *Clinical and Translational Science*. 13(5), pp. 848–860. doi:10.1111/cts.12769.
Ramírez, B., Niño-Orrego, M.J., Cárdenas, D., et al. (2019) Copy number variation profiling in pharmacogenetics CYP-450 and GST genes in Colombian population. *BMC Medical Genomics*. 12(1), pp. 110. doi:10.1186/s12920-019-0556-x.
Ramsay, M., Crowther, N., Tambo, E., et al. (2016) H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Global Health, Epidemiology and Genomics*. 1(20), pp. 1–13. doi:10.1017/gheg.2016.17.
Rao, Y., Hoffmann, E., Zia, M., et al. (2000) Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. *Molecular Pharmacology*. 58(4), pp. 747–755. doi:10.1124/mol.58.4.747.
Rausch, T., Zichner, T., Schlattl, A., et al. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 28(18), pp. 333–339. doi:10.1093/bioinformatics/bts378.
Rautio, J., Meanwell, N.A., Di, L., et al. (2018) The expanding role of prodrugs in contemporary drug design and development. *Nature Reviews Drug Discovery*. 17(8), pp. 559–587. doi:10.1038/nrd.2018.46.
Redon, R., Ishikawa, S., Fitch, K.R., et al. (2006) Global variation in copy number in the human genome. *Nature*. 444(7118), pp. 444–454. doi:10.1038/nature05329.
Rigau, M., Juan, D., Valencia, A., et al. (2019) Intronic CNVs and gene expression variation in human populations. *PLoS Genetics*. 15(1), pp. 1–23. doi:10.1371/journal.pgen.1007902.
Romdhane, L., Mezzi, N., Dallali, H., et al. (2021) A map of copy number variations in the Tunisian population: a valuable tool for medical genomics in North Africa. *NPJ Genomic Medicine*. 6(1), pp. 3. doi:10.1038/s41525-020-00166-5.
Rotimi, C., Abayomi, A., Abimiku, A., et al. (2014) Research capacity. Enabling the genomic revolution in Africa. *Science*. 344(6190), pp. 1346–1348. doi:10.1126/science.1251546.
Saif, M.W., Ezzeldin, H., Vance, K., et al. (2007) DPYD*2A mutation: the most common mutation associated with DPD deficiency. *Cancer Chemotherapy and Pharmacology*. 60(4), pp. 503–507. doi:10.1007/s00280-006-0392-5.
Saif, M.W., Ezzeldin, H., Vance, K., et al. (2019) Prevalence and properties of

intragenic copy-number variation in Mendelian disease genes. *Nucleic Acids Research*. 123(1), pp. 333–342. doi:10.1038/s41436-018-0033-5.
Saini, R., Saini, S. and Sugandha, R. (2010) Pharmacogenetics: The future medicine. *Journal of Advanced Pharmaceutical Technology & Research*. 1(4), pp. 423–424. doi:10.4103/0110-5558.76443.
Santos, M., Niemi, M., Hiratsuka, M., et al. (2018) Novel copy-number variations in pharmacogenes contribute to interindividual differences in drug pharmacokinetics. *Genetics in Medicine*. 20(6), pp. 622–629. doi:10.1038/gim.2017.156.
Sarkar, S. and Ganguly, A. (2016) Current Overview of Anti-Tuberculosis Drugs: Metabolism and Toxicities. *Mycobacterial Diseases*. 6(2), pp. 1–6. doi:10.4172/2161-1068.1000209.
Scherrmann, J.M. (2009) Transporters in absorption, distribution, and elimination. *Chemistry & Biodiversity*. 6(11), pp. 1933–1942. doi:10.1002/cbdv.200900171.
Schipani, A., Wyen, C., Mahungu, T., et al. (2011) Integration of population pharmacokinetics and pharmacogenetics: an aid to optimal nevirapine dose selection in HIV-infected individuals. *Journal of Antimicrobial Chemotherapy*. 66(6), pp. 1332–1339. doi:10.1093/jac/dkr087.
Scripture, C.D. and Pieper, J.A. (2001) Clinical pharmacokinetics of fluvastatin. *Clinical pharmacokinetics*. 40(4), pp. 263–281. doi:10.2165/00003088-200140040-00003.
Sedlazeck, F.J., Lee, H., Darby, C.A., et al. (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*. 19(6), pp. 329–346. doi:10.1038/s41576-018-0003-4.
Sedlazeck, F.J., Rescheneder, P., Smolka, M., et al. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*. 15(6), pp. 461–468. doi:10.1038/s41592-018-0001-7.
Semakula, J., Kisa, G., Mouton, H., et al. (2021) Anticoagulation in sub-Saharan Africa: Are Direct Oral Anticoagulants the answer? A review of lessons learnt from warfarin. *British Journal of Clinical Pharmacology*. 87(10), pp. 3699-3705. doi:10.1111/bcp.14796.
Semakula, J.R., Mouton, J.P., Jorgensen, A., et al. (2020) A cross-sectional evaluation of five warfarin anticoagulation services in Uganda and South Africa. *PLoS One*. 15(1), pp. 1–9. doi:10.1371/journal.pone.0227458.
Shaikh, T.H., Gai, X., Perin, J.C., et al. (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Research*. 19(9), pp. 1682–1690. doi:10.1101/gr.083501.108.
Sham, P.C. and Purcell, S.M. (2014) Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*. 15(5), pp. 335–346. doi:10.1038/nrg3706.
Sim, S.C., Kacevska, M. and Ingelman-Sundberg, M. (2013) Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. *The Pharmacogenomics Journal*. 13(1), pp. 1–11. doi:10.1038/tpj.2012.45.
Sistonen, J., Sajantila, A., Lao, O., et al. (2007) CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. *Pharmacogenetics and Genomics*. 17(2), pp. 93–101. doi:10.1097/01.fpc.0000239974.69464.f2.
Siwada, A. (2020) 'Identifying SNP haplotypes that tag Copy Number Variation in ADME genes in Sub-Saharan African population', *Honours report*, University of the Witwatersrand, Johannesburg.
Sloan, D.J., McCallum, A.D., Schipani, A., et al. (2017) Genetic Determinants of the Pharmacokinetic Variability of Rifampin in Malawian Adults with Pulmonary

Tuberculosis. *Antimicrobial Agents and Chemotherapy*. 61(7), pp. 1–9. doi:10.1128/AAC.00210-17.

de Smith, A.J., Tsalenko, A., Sampas, N., et al. (2007) Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Human Molecular Genetics*. 16(23), pp. 2783–2794. doi:10.1093/hmg/ddm208.

de Smith, A.J., Walters, R.G., Froguel, P., et al. (2009) Human genes involved in copy number variation: Mechanisms of origin, functional effects and implications for disease. *Cytogenetic and Genome Research*. 123(4), pp. 17–26. doi:10.1159/000184688.

Soko, N.D., Masimirembwa, C. and Dandara, C. (2018) Rosuvastatin pharmacogenetics in African populations. *Pharmacogenomics*. 19(18), pp. 1373–1375. doi:10.2217/pgs-2018-0168.

Spratlin, J. and Sawyer, M.B. (2007) Pharmacogenetics of paclitaxel metabolism. *Critical Reviews in Oncology/Hematology*. 61(3), pp. 222–229. doi:https://doi.org/10.1016/j.critrevonc.2006.09.006.

Stancu, M.C., van Roosmalen, M.J., Renkens, I., et al. (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications*. 8(1), pp. 1326. doi:10.1038/s41467-017-01343-4.

Steen, V.M., Andreassen, O.A., Daly, A.K., et al. (1995) Detection of the poor metabolizer-associated CYP2D6(D) gene deletion allele by long-PCR technology. *Pharmacogenetics*. 5(4), pp. 215–223. doi:10.1097/00008571-199508000-00005.

Stingl, J.C., Brockmöller, J. and Viviani, R. (2013) Genetic variability of drug-metabolizing enzymes: the dual impact on psychiatric therapy and regulation of brain function. *Molecular Psychiatry*. 18(3), pp. 273–287. doi:10.1038/mp.2012.42.

Stranger, B.E., Forrest, M.S., Dunning, M., et al. (2007) Relative impact of Nucleotide and Copy Number Variation on Gene Expression phenotypes. *Science*. 315(5813), pp. 848–853. doi:10.1126/science.1136678.

Takahashi, H. and Echizen, H. (2003) Pharmacogenetics of CYP2C9 and interindividual variability in anticoagulant response to warfarin. *The Pharmacogenomics Journal*. 3(4), pp. 202–214. doi:10.1038/sj.tpj.6500182.

Tattini, L., D'Aurizio, R. and Magi, A. (2015) Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*. 3, pp. 1–8. doi:10.3389/fbioe.2015.00092.

Teo, S.M., Pawitan, Y., Ku, C.S., et al. (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*. 28(21), pp. 2711–2718. doi:10.1093/bioinformatics/bts535.

The Cancer Atlas (2018) *Sub-Saharan Africa*. Available at: https://canceratlas.cancer.org/the-burden/sub-saharan-africa/ (Accessed: 2 May 2021).

del Tredici, A.L., Malhotra, A., Dedek, M., et al. (2018) Frequency of CYP2D6 Alleles Including Structural Variants in the United States. *Frontiers in Pharmacology*. 9, pp. 305. doi:10.3389/fphar.2018.00305.

Tremmel, R., Klein, K., Battke, F., et al. (2020) Copy number variation profiling in pharmacogenes using panel-based exome resequencing and correlation to human liver expression. *Human Genetics*. 139(2), pp. 137–149. doi:10.1007/s00439-019-02093-7.

Tremmel, R., Klein, K., Winter, S., et al. (2016) Gene copy number variation analysis reveals dosage-insensitive expression of CYP2E1. *The Pharmacogenomics Journal*. 16(6), pp. 551–558. doi:10.1038/tpj.2015.69.

Tyndale, R.F. and Sellers, E.M. (2001) Variable CYP2A6-mediated nicotine metabolism alters smoking behavior and risk. *Drug Metabolism and Disposition*. 29(4), pp. 548–552.

Ugartondo, N., Martínez-Gil, N., Esteve, M., et al. (2021) Functional Analyses of Four CYP1A1 Missense Mutations Present in Patients with Atypical Femoral Fractures. *International Journal of Molecular Sciences*. 22(14), pp. 1–11. doi:10.3390/ijms22147395.

Vijzelaar, R., Botton, M.R., Stolk, L., et al. (2018) Multi-ethnic SULT1A1 copy number profiling with multiplex ligation-dependent probe amplification. *Pharmacogenomics.* 19(9), pp. 761–770. doi:10.2217/pgs-2018-0047.

Wang, H., Veldink, J.H., Blauw, H., et al. (2009) Markov Models for inferring copy number variations from genotype data on Illumina platforms. *Human Heredity*. 68(1), pp. 1–22. doi:10.1159/000210445.

Wang, K., Li, M., Hadley, D., et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*. 17(11), pp. 1665–1674. doi:10.1101/gr.6861907.

Wang, P., Pradhan, K., Zhong, X.-B., et al. (2016) Isoniazid metabolism and hepatotoxicity. *Acta Pharmaceutica Sinica B*. 6(5), pp. 384–392. doi:10.1016/j.apsb.2016.07.014.

Warnich, L., Drögemöller, B.I., Pepper, M.S., et al. (2011) Pharmacogenomic Research in South Africa: Lessons Learned and Future Opportunities in the Rainbow Nation. *Current Pharmacogenomics and Personalized Medicine*. 9(3), pp. 191–207. doi:10.2174/187569211796957575.

Wasunna, B., Zurovac, D., Goodman, C.A., et al. (2008) Why don't health workers prescribe ACT? A qualitative study of factors affecting the prescription of artemether-lumefantrine. *Malaria journal*. 7, pp. 29. doi:10.1186/1475-2875-7-29.

Weinshilboum, R. (2003) Inheritance and Drug Response. *New England Journal of Medicine*. 348(6), pp. 529–537. doi:10.1056/NEJMra020021.

Wenger, A.M., Peluso, P., Rowell, W.J., et al. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*. 37(10), pp. 1155–1162. doi:10.1038/s41587-019-0217-9.

Wester, K., Jönsson, A., Spigset, O., et al. (2008) Incidence of fatal adverse drug reactions: A population based study. *British Journal of Clinical Pharmacology*. 65, pp. 573–579. doi:10.1111/j.1365-2125.2007.03064.x.

Wilkinson, G.R. (2005) Drug metabolism and variability among patients in drug response. *The New England Journal of Medicine*. 352(21), pp. 2211–2221. doi:10.1056/NEJMra032424.

Williams, J.A., Hyland, R., Jones, B.C., et al. (2004) Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metabolism and Disposition*. 32(11), pp. 1201–1208. doi:10.1124/dmd.104.000794.

Williamson, B., Dooley, K.E., Zhang, Y., et al. (2013) Induction of Influx and Efflux Transporters and Cytochrome P450 3A4 in Primary Human Hepatocytes by Rifampin, Rifabutin, and Rifapentine. *Antimicrobial Agents and Chemotherapy*. 57(12), pp. 6366–6369. doi:10.1128/AAC.01124-13.

Wilson, W., de Villena, F.P.M, Lyn-Cook, B.D., et al. (2004) Characterization of a common deletion polymorphism of the UGT2B17 gene linked to UGT2B15. *Genomics*. 84(4), pp. 707–714. doi:https://doi.org/10.1016/j.ygeno.2004.06.011.

World Health Organization (2018) *Updated recommendations on first-line and second-line antiretroviral regimens and post-exposure prophylaxis and recommendations*

on early infant diagnosis of HIV. Available at: https://www.who.int/publications/i/item/WHO-CDS-HIV-18.51 (Accessed: 20 April 2020).
World Health Organization (2019) *The 'World malaria report 2019' at a glance.* Available at: https://www.who.int/news-room/feature-stories/detail/world-malaria-report-2019 (Accessed: 30 April 2020).
World Health Organization (2020a) *Malaria*. Available at: https://www.who.int/news-room/fact-sheets/detail/malaria (Accessed: 30 April 2020).
World Health Organization (2020b) *Tuberculosis*. Available at: https://www.who.int/health-topics/tuberculosis#tab=tab_1 (Accessed: 30 April 2020).
Yan, M.M., Wu, S.S., Ying, Y.Q., et al. (2018) Safety assessment of concurrent statin treatment and evaluation of drug interactions in China. *SAGE Open Medicine*. 6, pp. 1–9. doi:10.1177/2050312118798278.
Yao, R., Yu, T., Qing, Y., *et al*. (2019) Evaluation of copy number detection from panel-based next-generation sequencing data. *Molecular Genetics and Genomic Medicine.* 7(1), pp. 1–8. doi: 10.1002/mgg3.513.
Ye, K., Schulz, M.H., Long, Q., et al. (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 25(21), pp. 2865–2871. doi:10.1093/bioinformatics/btp394.
Yoon, S., Xuan, Z., Makarov, V., et al. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*. 19(9), pp. 1586–1592. doi:10.1101/gr.092981.109.
Yu, X., Kubota, T., Dhakal, I., et al. (2013) Copy number variation in sulfotransferase isoform 1A1 (SULT1A1) is significantly associated with enzymatic activity in Japanese subjects. *Pharmacogenomics and Personalized Medicine*. 6, pp. 19–24. doi:10.2147/PGPM.S36579.
Zang, M., Zhu, F., Zhao, L., et al. (2014) The effect of UGTs polymorphism on the auto-induction phase II metabolism-mediated pharmacokinetics of dihydroartemisinin in healthy Chinese subjects after oral administration of a fixed combination of dihydroartemisinin-piperaquine. *Malaria Journal*. 13(1), pp. 478. doi:10.1186/1475-2875-13-478.
Zanger, U.M. and Klein, K. (2013) Pharmacogenetics of cytochrome P450 2B6 (CYP2B6): Advances on polymorphisms, mechanisms, and clinical relevance. *Frontiers in Genetics*. 4(24), pp. 1–12. doi:10.3389/fgene.2013.00024.
Zhang, J., Wang, J. and Wu, Y. (2012) An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinformatics*. 13(6), pp. 1–11. doi: 10.1186/1471-2105-13-S6-S6.
Zhang, L., Bai, W., Yuan, N., et al. (2019) Comprehensively benchmarking applications for detecting copy number variation. *PLoS Computational Biology*. 15(5), pp. 1–12. doi:10.1371/journal.pcbi.1007069.
Zhang, L., Shi, J., Ouyang, J., et al. (2021) X-CNV: genome-wide prediction of the pathogenicity of copy number variations. *Genome Medicine*. 13(1), pp. 132. doi:10.1186/s13073-021-00945-4.
Zhang, Z., Cheng, H., Hong, X., et al. (2019) EnsembleCNV: an ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data. *Nucleic Acids Research*. 47(7), pp. 1–13. doi:10.1093/nar/gkz068.
Zhang, Z., Lange, K. and Sabatti, C. (2012) Reconstructing DNA copy number by joint segmentation of multiple sequences. *BMC bioinformatics*. 13, pp. 205. doi:10.1186/1471-2105-13-205.
Zhang, Z.D., Du, J., Lam, H., et al. (2011) Identification of genomic indels and structural variations using split reads. *BMC genomics*. 12, pp. 375. doi:10.1186/1471-2164-

12-375.
Zhao, M., Wang, Qingguo, Wang, Quan, et al. (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 14(11), pp. 1–16. doi:10.1186/1471-2105-14-S11-S1.
Zhao, X., Collins, R.L., Lee, W.-P., et al. (2021) Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *The American Journal of Human Genetics*. 108(5), pp. 919–928. doi:https://doi.org/10.1016/j.ajhg.2021.03.014.
Zhou, B., Ho, S.S., Zhang, X., et al. (2017) Whole-genome sequencing analysis of genomic copy number variation (CNV) using low-coverage and paired-end strategies is highly efficient and outperforms array based CNV analysis. *Journal of Medical Genetics*. 55(11), pp. 735–743. doi: 10.1136/jmedgenet-2018-105272.
Zhou, K. and Pearson, E. (2013). Insights from genome-wide association studies of drug response. *Annual Review of pharmacology and toxicology*. 53, pp. 299–310. doi: 10.1146/annurev-pharmtox-011112-140237.
Zhou, S.F. (2009) Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I. *Clinical Pharmacokinetics*. 48(11), pp. 689–723. doi:10.2165/11318030-000000000-00000.
Zhou, X.Y., Hu, X.X., Wang, C.C., et al. (2019) Enzymatic Activities of CYP3A4 Allelic Variants on Quinine 3-Hydroxylation In Vitro. *Frontiers in Pharmacology*. 10, pp. 591. doi:10.3389/fphar.2019.00591.
Zhou, Z., Wang, W., Wang, L.S., et al. (2018) Integrative DNA copy number detection and genotyping from sequencing and array-based platforms. *Bioinformatics*. 34(14), pp. 2349–2355. doi:10.1093/bioinformatics/bty104.
Zhu, H. and Zhou, X. (2020) Statistical methods for SNP heritability estimation and partition: A review. *Computational and Structural Biotechnology Journal*. 18, pp. 1557–1568. doi: 10.1016/j.csbj.2020.06.011.

7. APPENDICES

APPENDIX A: PharmaADME core and extended ADME gene lists.

Table A1: Core ADME gene list (PharmaADME.org, 2021).

Gene Symbol	Full Gene Name	Class
ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	Transporter
ABCC2	ATP-binding cassette, sub-family C (CFTR/MRP), member 2	Transporter
ABCG2	ATP-binding cassette, sub-family G (WHITE), member 2	Transporter
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1	Phase I
CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2	Phase I
CYP2A6	cytochrome P450, family 2, subfamily A, polypeptide 6	Phase I
CYP2B6	cytochrome P450, family 2, subfamily B, polypeptide 6	Phase I
CYP2C19	cytochrome P450, family 2, subfamily C, polypeptide 19	Phase I
CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8	Phase I
CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	Phase I
CYP2D6	cytochrome P450, family 2, subfamily D, polypeptide 6	Phase I
CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1	Phase I
CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	Phase I
CYP3A5	cytochrome P450, family 3, subfamily A, polypeptide 5	Phase I
DPYD	dihydropyrimidine dehydrogenase	Phase I
GSTM1	glutathione S-transferase M1	Phase II
GSTP1	glutathione S-transferase pi	Phase II
GSTT1	glutathione S-transferase theta 1	Phase II
NAT1	N-acetyltransferase 1 (arylamine N-acetyltransferase)	Phase II
NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)	Phase II
SLC15A2	solute carrier family 15 (H+/peptide transporter), member 2	Transporter
SLC22A1	solute carrier family 22 (organic cation transporter), member 1	Transporter
SLC22A2	solute carrier family 22 (organic cation transporter), member 2	Transporter
SLC22A6	solute carrier family 22 (organic anion transporter), member 6	Transporter
SLCO1B1	solute carrier organic anion transporter family, member 1B1	Transporter
SLCO1B3	solute carrier organic anion transporter family, member 1B3	Transporter
SULT1A1	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	Phase II
TPMT	thiopurine S-methyltransferase,	Phase II
UGT1A1	UDP glucuronosyltransferase 1 family, polypeptide A1	Phase II
UGT2B15	UDP glucuronosyltransferase 2 family, polypeptide B15	Phase II
UGT2B17	UDP glucuronosyltransferase 2 family, polypeptide B17	Phase II
UGT2B7	UDP glucuronosyltransferase 2 family, polypeptide B7	Phase II

Table A2: Extended ADME gene list. The rankings of the extended ADME genes have been established based on input from seven major pharmaceutical companies in line with their interest in seeing a particular gene on a product (PharmaADME.org, 2021).

Rank	Gene Symbol	Full Gene Name	Class
7	*ABCB8*	ATP-binding cassette, sub-family B (MDR/TAP), member 8	Transporter
7	*ABCC12*	ATP-binding cassette, sub-family C (CFTR/MRP), member 12	Transporter
7	*ABCC3*	ATP-binding cassette, sub-family C (CFTR/MRP), member 3	Transporter
7	*ABCC4*	ATP-binding cassette, sub-family C (CFTR/MRP), member 4	Transporter
7	*AHR*	aryl hydrocarbon receptor	Modifier
7	*ALDH4A1*	aldehyde dehydrogenase 4 family, member A1	Phase I
7	*ALDH5A1*	aldehyde dehydrogenase 5 family, member A1	Phase I
7	*ALDH6A1*	aldehyde dehydrogenase 6 family, member A1	Phase I
7	*CES1*	carboxylesterase 1 (monocyte/macrophage serine esterase 1)	Phase I
7	*CES2*	carboxylesterase 2 (intestine, liver)	Phase I
7	*CYP7A1*	cytochrome P450, family 7, subfamily A, polypeptide 1	Phase I
7	*EPHX1*	epoxide hydrolase 1, microsomal (xenobiotic)	Phase I
7	*FMO3*	flavin containing monooxygenase 3	Phase I
7	*GSTA1*	glutathione S-transferase A1	Phase II
7	*GSTA2*	glutathione S-transferase A2	Phase II
7	*GSTA3*	glutathione S-transferase A3	Phase II
7	*GSTA4*	glutathione S-transferase A4	Phase II
7	*GSTA5*	glutathione S-transferase A5	Phase II
7	*GSTM2*	glutathione S-transferase M2 (muscle),glutathione S-transferase M4	Phase II
7	*GSTM3*	glutathione S-transferase M3 (brain)	Phase II
7	*GSTM4*	glutathione S-transferase M4	Phase II
7	*GSTO1*	glutathione S-transferase omega 1,glutathione S-transferase omega 2	Phase II
7	*GSTO2*	glutathione S-transferase omega 2	Phase II
7	*GSTT2*	glutathione S-transferase theta 2	Phase II
7	*SLC10A1*	solute carrier family 10 (sodium/bile acid cotransporter family), member 1	Transporter
7	*SLC15A1*	solute carrier family 15 (oligopeptide transporter), member 1	Transporter
7	*SLC22A11*	solute carrier family 22 (organic anion/cation transporter), member 11	Transporter
7	*SLC22A8*	solute carrier family 22 (organic anion transporter), member 8	Transporter
7	*SLC7A5*	solute carrier family 7 (cationic amino acid transporter, y+ system), member 5	Transporter
7	*SLCO1A2*	solute carrier organic anion transporter family, member 1A2	Transporter
7	*SLCO2B1*	solute carrier organic anion transporter family, member 2B1	Transporter
7	*SULT1A2*	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 2	Phase II
7	*SULT1A3*	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 3	Phase II
7	*SULT1B1*	sulfotransferase family, cytosolic, 1B, member 1	Phase II
7	*UGT1A3*	UDP glucuronosyltransferase 1 family, polypeptide A3	Phase II
7	*UGT1A6*	UDP glucuronosyltransferase 1 family, polypeptide A6	Phase II
7	*UGT1A7*	UDP glucuronosyltransferase 1 family, polypeptide A7	Phase II

7	*UGT1A8*	UDP glucuronosyltransferase 1 family, polypeptide A8	Phase II
7	*UGT1A9*	UDP glucuronosyltransferase 1 family, polypeptide A9	Phase II
7	*UGT2A1*	UDP glucuronosyltransferase 2 family, polypeptide A1	Phase II
7	*UGT2B11*	UDP glucuronosyltransferase 2 family, polypeptide B11	Phase II
7	*UGT2B28*	UDP glucuronosyltransferase 2 family, polypeptide B28	Phase II
7	*UGT2B4*	UDP glucuronosyltransferase 2 family, polypeptide B4	Phase II
6	*ABCA1*	ATP-binding cassette, sub-family A (ABC1), member 1	Transporter
6	*ABCA4*	ATP-binding cassette, sub-family A (ABC1), member 4	Transporter
6	*ABCB11*	ATP-binding cassette, sub-family B (MDR/TAP), member 11	Transporter
6	*ABCB4*	ATP-binding cassette, sub-family B (MDR/TAP), member 4	Transporter
6	*ABCB5*	ATP-binding cassette, sub-family B (MDR/TAP), member 5	Transporter
6	*ABCB6*	ATP-binding cassette, sub-family B (MDR/TAP), member 6	Transporter
6	*ABCB7*	ATP-binding cassette, sub-family B (MDR/TAP), member 7	Transporter
6	*ABCC1*	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	Transporter
6	*ABCC10*	ATP-binding cassette, sub-family C (CFTR/MRP), member 10	Transporter
6	*ABCC11*	ATP-binding cassette, sub-family C (CFTR/MRP), member 11	Transporter
6	*ABCC5*	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	Transporter
6	*ABCC6*	ATP-binding cassette, sub-family C (CFTR/MRP), member 6	Transporter
6	*ABCC8*	ATP-binding cassette, sub-family C (CFTR/MRP), member 8	Transporter
6	*ABCC9*	ATP-binding cassette, sub-family C (CFTR/MRP), member 9	Transporter
6	*ABCG1*	ATP-binding cassette, sub-family G (WHITE), member 1	Transporter
6	*ADH1A*	alcohol dehydrogenase 1A (class I), alpha polypeptide	Phase I
6	*ADH1B*	alcohol dehydrogenase IB (class I), beta polypeptide	Phase I
6	*ADH1C*	alcohol dehydrogenase 1C (class I), gamma polypeptide	Phase I
6	*ADH4*	alcohol dehydrogenase 4 (class II), pi polypeptide	Phase I
6	*ADH5*	alcohol dehydrogenase 5 (class III), chi polypeptide,methionyl aminopeptidase 1	Phase I
6	*ADH6*	alcohol dehydrogenase 6 (class V)	Phase I
6	*ADH7*	alcohol dehydrogenase 7 (class IV), mu or sigma polypeptide	Phase I
6	*ALDH1A1*	aldehyde dehydrogenase 1 family, member A1	Phase I
6	*ALDH1A2*	aldehyde dehydrogenase 1 family, member A2	Phase I
6	*ALDH1A3*	aldehyde dehydrogenase 1 family, member A3	Phase I
6	*ALDH1B1*	aldehyde dehydrogenase 1 family, member B1	Phase I
6	*ALDH2*	aldehyde dehydrogenase 2 family (mitochondrial)	Phase I
6	*ALDH3A1*	aldehyde dehydrogenase 3 family, memberA1	Phase I
6	*ALDH3A2*	aldehyde dehydrogenase 3 family, member A2	Phase I
6	*ALDH3B1*	aldehyde dehydrogenase 3 family, member B1	Phase I
6	*ALDH3B2*	aldehyde dehydrogenase 3 family, member B2	Phase I
6	*ALDH7A1*	aldehyde dehydrogenase 7 family, member A1	Phase I
6	*ALDH8A1*	aldehyde dehydrogenase 8 family, member A1	Phase I
6	*ALDH9A1*	aldehyde dehydrogenase 9 family, member A1	Phase I
6	*AOX1*	aldehyde oxidase 1	Phase I
6	*ARNT*	aryl hydrocarbon receptor nuclear translocator	Modifier
6	*CBR1*	carbonyl reductase 1	Phase I

6	*CBR3*	carbonyl reductase 3	Phase I
6	*CDA*	cytidine deaminase	Modifier
6	*CYB5R3*	cytochrome b5 reductase 3	Phase I
6	*CYP11A1*	cytochrome P450, family 11, subfamily A, polypeptide 1	Phase I
6	*CYP11B1*	cytochrome P450, family 11, subfamily B, polypeptide 1	Phase I
6	*CYP11B2*	cytochrome P450, family 11, subfamily B, polypeptide 2	Phase I
6	*CYP17A1*	cytochrome P450, family 17, subfamily A, polypeptide 1	Phase I
6	*CYP1B1*	cytochrome P450, family 1, subfamily B, polypeptide 1	Phase I
6	*CYP20A1*	cytochrome P450, family 20, subfamily A, polypeptide 1	Phase I
6	*CYP20A1*	cytochrome P450, family 20, subfamily A, polypeptide 1	Phase I
6	*CYP21A2*	cytochrome P450, family 21, subfamily A, polypeptide 2	Phase I
6	*CYP24A1*	cytochrome P450, family 24, subfamily A, polypeptide 1	Phase I
6	*CYP26A1*	cytochrome P450, family 26, subfamily A, polypeptide 1	Phase I
6	*CYP27A1*	cytochrome P450, family 27, subfamily A, polypeptide 1	Phase I
6	*CYP2A13*	cytochrome P450, family 2, subfamily A, polypeptide 13	Phase I
6	*CYP2A7*	cytochrome P450, family 2, subfamily A, polypeptide 7	Phase I
6	*CYP2C18*	cytochrome P450, family 2, subfamily C, polypeptide 18	Phase I
6	*CYP2F1*	cytochrome P450, family 2, subfamily F, polypeptide 1	Phase I
6	*CYP2J2*	cytochrome P450, family 2, subfamily J, polypeptide 2	Phase I
6	*CYP39A1*	cytochrome P450, family 39, subfamily A, polypeptide 1	Phase I
6	*CYP3A43*	cytochrome P450, family 3, subfamily A, polypeptide 43	Phase I
6	*CYP3A7*	cytochrome P450, family 3, subfamily A, polypeptide 7	Phase I
6	*CYP4B1*	cytochrome P450, family 4, subfamily B, polypeptide 1	Phase I
6	*CYP4F11*	cytochrome P450, family 4, subfamily F, polypeptide 11	Phase I
6	*CYP51A1*	cytochrome P450, family 51, subfamily A, polypeptide 1	Phase I
6	*EPHX2*	epoxide hydrolase 2, cytoplasmic	Phase I
6	*FMO1*	flavin containing monooxygenase 1	Phase I
6	*FMO2*	flavin containing monooxygenase 2	Phase I
6	*FMO4*	flavin containing monooxygenase 4	Phase I
6	*FMO5*	flavin containing monooxygenase 5	Phase I
6	*GPX2*	glutathione peroxidase 2 (gastrointestinal)	Phase I
6	*GPX3*	glutathione peroxidase 3 (plasma)	Phase I
6	*GPX7*	glutathione peroxidase 7	Phase I
6	*GSR*	glutathione reductase	Phase I
6	*GSTK1*	glutathione S-transferase kappa 1	Phase II
6	*GSTM5*	glutathione S-transferase M5	Phase II
6	*GSTZ1*	glutathione transferase zeta 1 (maleylacetoacetate isomerase)	Phase II
6	*NNMT*	nicotinamide N-methyltransferase	Phase II
6	*NR1I2*	nuclear receptor subfamily 1, group I, member 2	Modifier
6	*NR1I3*	nuclear receptor subfamily 1, group I, member 3	Modifier
6	*PNMT*	phenylethanolamine N-methyltransferase	Phase II
6	*PON1*	paraoxonase 1	Phase I
6	*PON2*	paraoxonase 2	Phase I
6	*PON3*	paraoxonase 3	Phase I

6	*POR*	P450 (cytochrome) oxidoreductase	Modifier
6	*PPARD*	peroxisome proliferative activated receptor, delta	Modifier
6	*PPARG*	peroxisome proliferative activated receptor, gamma	Modifier
6	*RXRA*	retinoid X receptor, alpha	Modifier
6	*SLC10A2*	solute carrier family 10 (sodium/bile acid cotransporter family), member 2	Transporter
6	*SLC13A1*	solute carrier family 13 (sodium/sulfate symporters), member 1	Transporter
6	*SLC13A2*	solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 2	Transporter
6	*SLC13A3*	solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 3	Transporter
6	*SLC16A1*	solute carrier family 16 (monocarboxylic acid Transporter), member 1	Transporter
6	*SLC19A1*	solute carrier family 19 (folate transporter), member 1	Transporter
6	*SLC22A10*	solute carrier family 22 (organic anion/cation transporter), member 10	Transporter
6	*SLC22A12*	solute carrier family 22 (organic anion/cation transporter), member 12	Transporter
6	*SLC22A13*	solute carrier family 22 (organic cation transporter), member 13	Transporter
6	*SLC22A14*	solute carrier family 22 (organic cation transporter), member 14	Transporter
6	*SLC22A15*	solute carrier family 22 (organic cation transporter), member 15	Transporter
6	*SLC22A16*	solute carrier family 22 (organic cation transporter), member 16	Transporter
6	*SLC22A17*	solute carrier family 22 (organic cation transporter), member 17	Transporter
6	*SLC22A18*	solute carrier family 22 (organic cation transporter), member 18	Transporter
6	*SLC22A18AS*	solute carrier family 22 (organic cation transporter), member 18 antisense	Transporter
6	*SLC22A3*	solute carrier family 22 (extraneuronal monoamine transporter), member 3	Transporter
6	*SLC22A4*	solute carrier family 22 (organic cation transporter), member 4	Transporter
6	*SLC22A5*	solute carrier family 22 (organic cation transporter), member 5	Transporter
6	*SLC22A7*	solute carrier family 22 (organic anion transporter), member 7	Transporter
6	*SLC22A9*	solute carrier family 22 (organic anion/cation transporter), member 9	Transporter
6	*SLC27A1*	solute carrier family 27 (fatty acid transporter), member 1	Transporter
6	*SLC28A1*	solute carrier family 28 (sodium-coupled nucleoside transporter), member 1	Transporter
6	*SLC28A2*	solute carrier family 28 (sodium-coupled nucleoside transporter), member 2	Transporter
6	*SLC28A3*	solute carrier family 28 (sodium-coupled nucleoside transporter), member 3	Transporter
6	*SLC29A1*	solute carrier family 29 (nucleoside Transporter), member 1	Transporter
6	*SLC29A2*	solute carrier family 29 (nucleoside Transporter), member 2	Transporter
6	*SLC2A4*	solute carrier family 2 (facilitated glucose transporter), member 4	Transporter
6	*SLC2A5*	solute carrier family 2 (facilitated glucose/fructose transporter), member 5	Transporter
6	*SLC5A6*	solute carrier family 5 (sodium-dependent vitamin transporter)	Transporter
6	*SLC6A6*	solute carrier family 6 (neurotransmitter transporter, taurine), member 6	Transporter
6	*SLC7A8*	solute carrier family 7 (cationic amino acid transporter, y+ system), member 8	Transporter
6	*SLCO1C1*	solute carrier organic anion transporter family, member 1C1	Transporter
6	*SLCO2A1*	solute carrier organic anion transporter family, member 2A1	Transporter
6	*SLCO3A1*	solute carrier organic anion transporter family, member 3A1	Transporter

6	*SLCO4A1*	solute carrier organic anion transporter family, member 4A1	Transporter
6	*SLCO4C1*	solute carrier organic anion transporter family, member 4C1	Transporter
6	*SLCO5A1*	solute carrier organic anion transporter family, member 5A1	Transporter
6	*SLCO6A1*	solute carrier organic anion transporter family, member 6A1	Transporter
6	*SULT1C1*	sulfotransferase family, cytosolic, 1C, member 1	Phase II
6	*SULT1C2*	sulfotransferase family, cytosolic, 1C, member 2	Phase II
6	*SULT1E1*	sulfotransferase family 1E, estrogen-preferring, member 1	Phase II
6	*SULT2A1*	sulfotransferase family, cytosolic, 2A, DHEA preferring, member 1	Phase II
6	*SULT2B1*	sulfotransferase family, cytosolic, 2B, member 1	Phase II
6	*TAP1*	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)	Transporter
6	*UGT1A10*	UDP glucuronosyltransferase 1 family, polypeptide A10	Phase II
6	*UGT1A4*	UDP glucuronosyltransferase 1 family, polypeptide A4	Phase II
6	*UGT1A5*	UDP glucuronosyltransferase 1 family, polypeptide A5	Phase II
6	*UGT2B10*	UDP glucuronosyltransferase 2 family, polypeptide B10	Phase II
5	*ABCC13*	ATP-binding cassette, sub-family C (CFTR/MRP), member 13	Transporter
5	*ARSA*	arylsulfatase A	Modifier
5	*CAT*	catalase	Modifier
5	*CHST8*	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 8	Phase II
5	*CYP19A1*	cytochrome P450, family 19, subfamily A, polypeptide 1	Phase I
5	*CYP26C1*	cytochrome P450, family 26, subfamily C, polypeptide 1	Phase I
5	*CYP27B1*	cytochrome P450, family 27, subfamily B, polypeptide 1	Phase I
5	*CYP2R1*	cytochrome P450, family 2, subfamily R, polypeptide 1	Phase I
5	*CYP2S1*	cytochrome P450, family 2, subfamily S, polypeptide 1	Phase I
5	*CYP46A1*	cytochrome P450, family 46, subfamily A, polypeptide 1	Phase I
5	*CYP4A11*	cytochrome P450, family 4, subfamily A, polypeptide 11	Phase I
5	*CYP4F12*	cytochrome P450, family 4, subfamily F, polypeptide 12	Phase I
5	*CYP4F2*	cytochrome P450, family 4, subfamily F, polypeptide 2	Phase I
5	*CYP4F3*	cytochrome P450, family 4, subfamily F, polypeptide 3	Phase I
5	*CYP4F8*	cytochrome P450, family 4, subfamily F, polypeptide 8	Phase I
5	*CYP4Z1*	cytochrome P450, family 4, subfamily Z, polypeptide 1	Phase I
5	*CYP7B1*	cytochrome P450, family 7, subfamily B, polypeptide 1	Phase I
5	*CYP8B1*	cytochrome P450, family 8, subfamily B, polypeptide 1	Phase I
5	*DHRS13*	dehydrogenase/reductase (SDR family) member 13	Phase I
5	*DHRS2*	dehydrogenase/reductase (SDR family) member 2	Phase I
5	*GPX1*	glutathione peroxidase 1	Phase I
5	*GPX4*	glutathione peroxidase 4 (phospholipid hydroperoxidase)	Phase I
5	*GPX5*	glutathione peroxidase 5 (epididymal androgen-related protein)	Phase I
5	*GPX6*	glutathione peroxidase 6 (olfactory)	Phase I
5	*GSS*	glutathione synthetase	Phase I
5	*GSTCD*	glutathione S-transferase, C-terminal domain containing	Phase II
5	*HNF4A*	hepatocyte nuclear factor 4, alpha	Modifier
5	*HNMT*	histamine N-methyltransferase	Phase II
5	*HSD11B1*	hydroxysteroid (17-beta) dehydrogenase 11	Phase I
5	*HSD17B11*	hydroxysteroid (17-beta) dehydrogenase 11	Phase I

5	*HSD17B14*	hydroxysteroid (17-beta) dehydrogenase 14	Phase I
5	*LOC731356*	similar to dehydrogenase/reductase (SDR family) member 4 like 2	Phase I
5	*MGST1*	microsomal glutathione S-transferase 1	Phase II
5	*MGST2*	microsomal glutathione S-transferase 2	Phase II
5	*MGST3*	microsomal glutathione S-transferase 3	Phase II
5	*MPO*	myeloperoxidase	Modifier
5	*NOS1*	nitric oxide synthase 1 (neuronal)	Phase I
5	*NOS2A*	nitric oxide synthase 2A (inducible, hepatocytes)	Phase I
5	*NOS3*	nitric oxide synthase 3 (endothelial cell)	Phase I
5	*PPARA*	peroxisome proliferator-activated receptor alpha	Modifier
5	*SERPINA7*	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 7	Modifier
5	*SLC7A7*	solute carrier family 7 (cationic amino acid transporter, y+ system), member 7	Transporter
5	*SOD1*	superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))	Modifier
5	*SOD2*	superoxide dismutase 2, mitochondrial	Modifier
5	*SOD3*	superoxide dismutase 3, extracellular precursor	Modifier
5	*SULF1*	sulfatase 1	Phase I
5	*SULT4A1*	sulfotransferase family 4A, member 1	Phase II
5	*TAP2*	transporter 2, ATP-binding cassette, sub-family B (MDR/TAP)	Transporter
5	*UGT8*	UDP glycosyltransferase 8 (UDP-galactose ceramide galactosyltransferase)	Phase II
5	*XDH*	xanthine dehydrogenase	Phase I
4	*ADHFE1*	alcohol dehydrogenase, iron containing, 1	Phase I
4	*CHST1*	carbohydrate (keratan sulfate Gal-6) sulfotransferase 1	Phase II
4	*CHST10*	carbohydrate sulfotransferase 10	Phase II
4	*CHST11*	carbohydrate (chondroitin 4) sulfotransferase 11	Phase II
4	*CHST12*	carbohydrate (chondroitin 4) sulfotransferase 12	Phase II
4	*CHST13*	carbohydrate (chondroitin 4) sulfotransferase 13	Phase II
4	*CHST2*	carbohydrate (N-acetylglucosamine-6-O) sulfotransferase 2	Phase II
4	*CHST3*	carbohydrate (chondroitin 6) sulfotransferase 3	Phase II
4	*CHST4*	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 4	Phase II
4	*CHST5*	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 5	Phase II
4	*CHST6*	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 6	Phase II
4	*CHST7*	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 7	Phase II
4	*CHST9*	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 9	Phase II
4	*CYP2D7P1*	cytochrome P450, family 2, subfamily D, polypeptide 7 pseudogene 1	Phase I
4	*DDO*	D-aspartate oxidase	Phase I
4	*DHRS1*	dehydrogenase/reductase (SDR family) member 1	Phase I
4	*DHRS12*	dehydrogenase/reductase (SDR family) member 12	Phase I
4	*DHRS3*	dehydrogenase/reductase (SDR family) member 3	Phase I
4	*DHRS4*	dehydrogenase/reductase (SDR family) member 4	Phase I
4	*DHRS4L1*	dehydrogenase/reductase (SDR family) member 4 like 1	Phase I
4	*DHRS4L2*	dehydrogenase/reductase (SDR family) member 4 like 2	Phase I

4	*DHRS7*	dehydrogenase/reductase (SDR family) member 7	Phase I
4	*DHRS7B*	dehydrogenase/reductase (SDR family) member 7B	Phase I
4	*DHRS7C*	dehydrogenase/reductase (SDR family) member 7C	Phase I
4	*DHRS9*	dehydrogenase/reductase (SDR family) member 9	Phase I
4	*DHRSX*	dehydrogenase/reductase (SDR family) X-linked	Phase I
4	*DPEP1*	dipeptidase 1 (renal)	Phase I
4	*FMO6P*	flavin containing monooxygenase 6	Phase I
4	*HAGH*	hydroxyacylglutathione hydrolase	Phase I
4	*IAPP*	islet amyloid polypeptide	Modifier
4	*KCNJ11*	potassium inwardly-rectifying channel, subfamily J, member 11	Modifier
4	*LOC728667*	similar to dehydrogenase/reductase (SDR family) member 2 isoform 1	Phase I
4	*LOC731931*	similar to dehydrogenase/reductase (SDR family) member 2 isoform 1	Phase I
4	*MAT1A*	methionine adenosyltransferase I, alpha	Modifier
4	*METAP1*	methionyl aminopeptidase 1	Phase I
4	*PDE3A*	phosphodiesterase 3A, cGMP-inhibited	Phase I
4	*PDE3B*	phosphodiesterase 3B, cGMP-inhibited	Phase I
4	*PLGLB1*	plasminogen-like B1	Phase I
3	*ATP7A*	ATPase, Cu++ transporting, alpha polypeptide (Menkes syndrome)	Modifier
3	*ATP7B*	ATPase, Cu++ transporting, beta polypeptide	Modifier
3	*CFTR*	cystic fibrosis transmembrane conductance regulator	Modifier

APPENDIX B: Depth of coverage plots for long-read sequences.

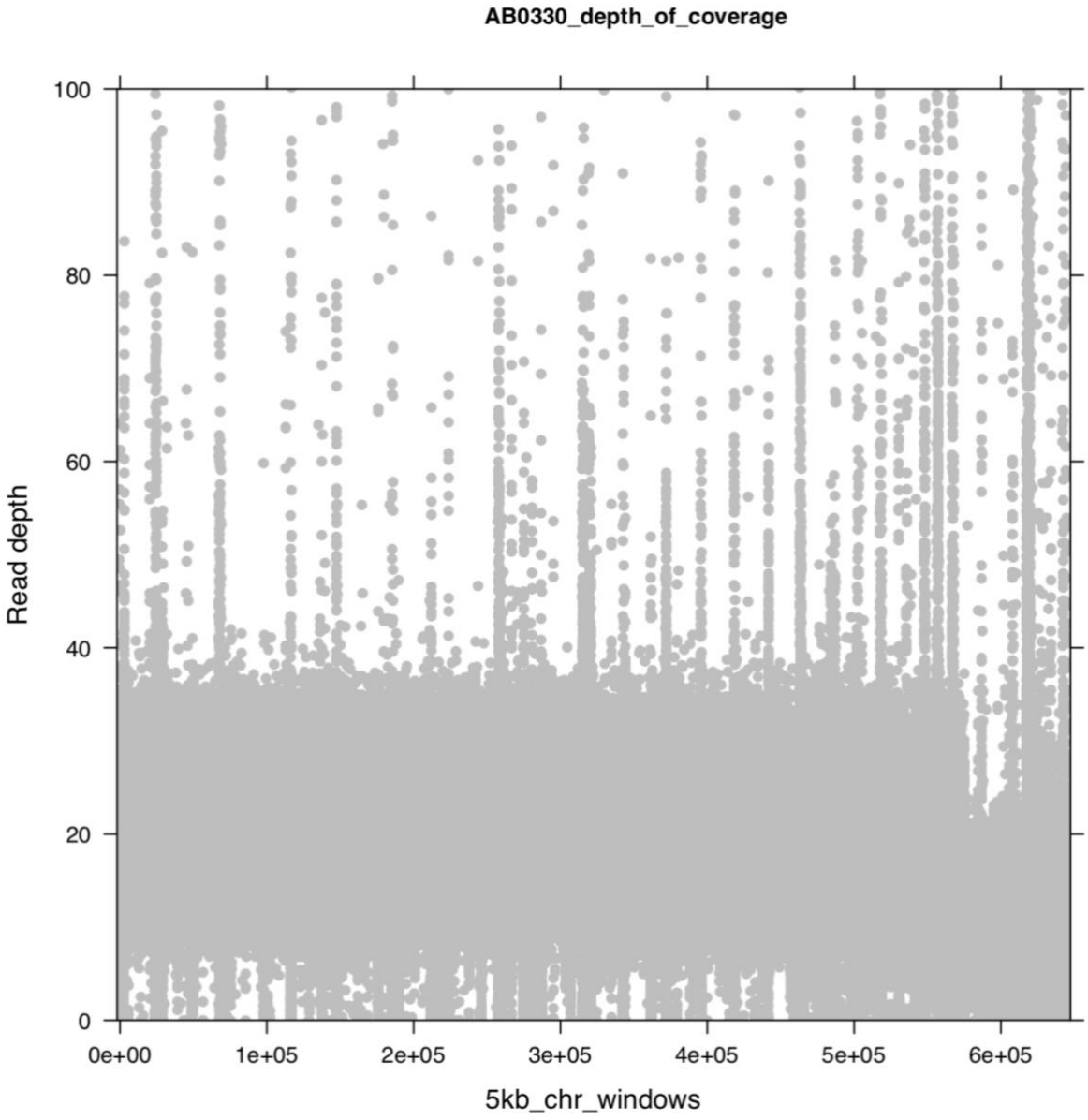


Figure B1: AB0330 depth of coverage.

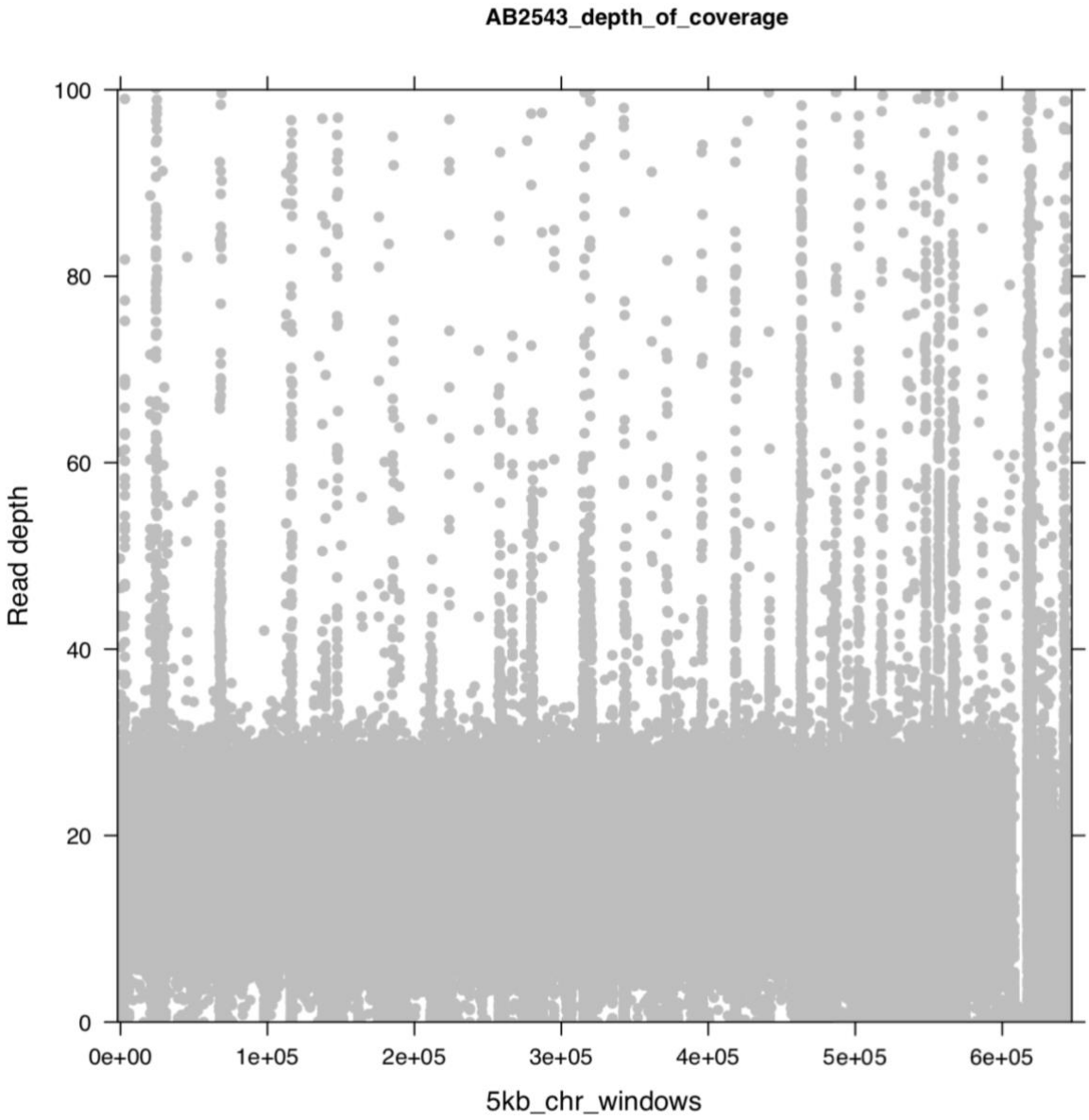


Figure B2: AB2543 depth of coverage.

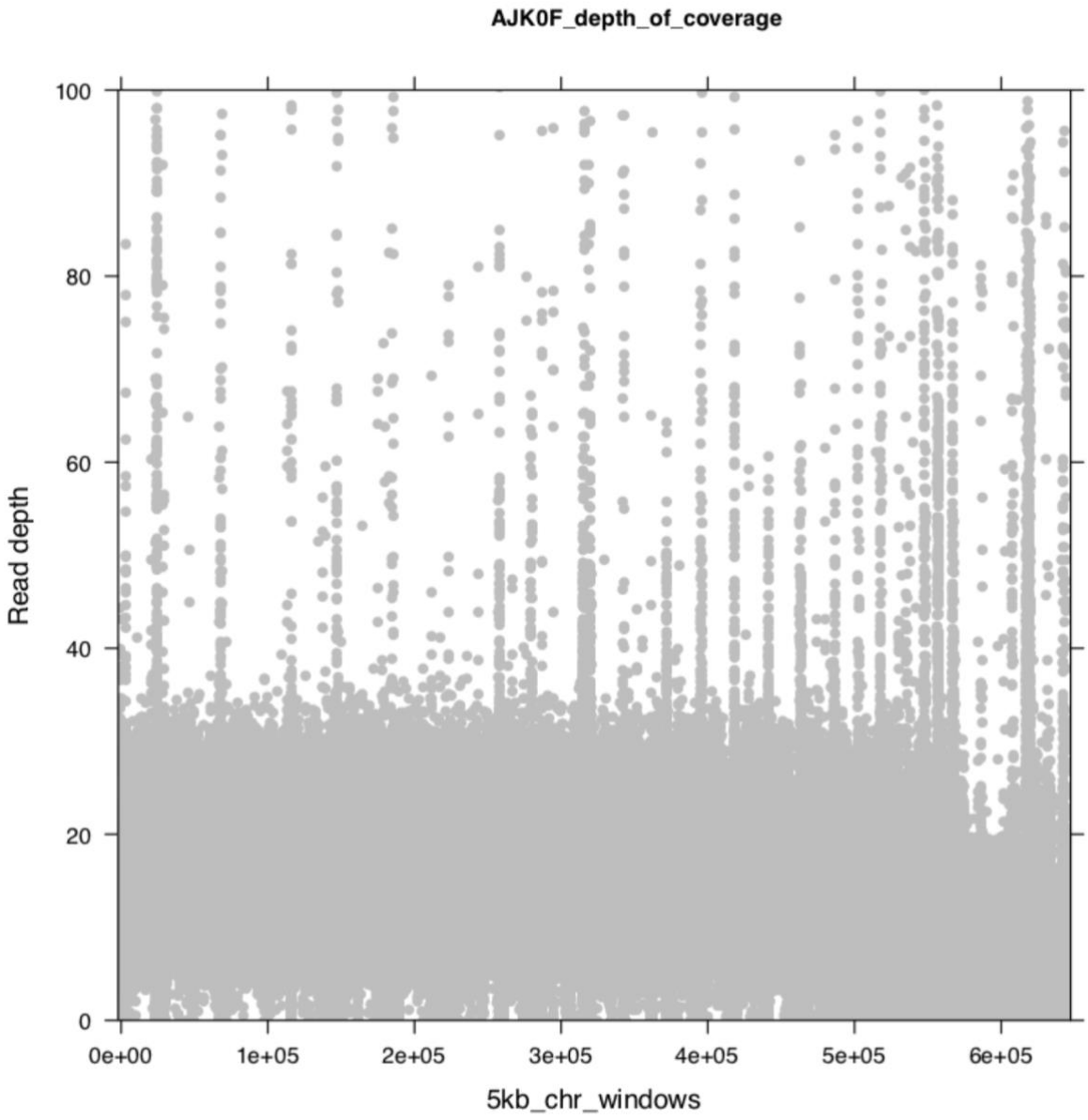


Figure B3: AJK0F depth of coverage.

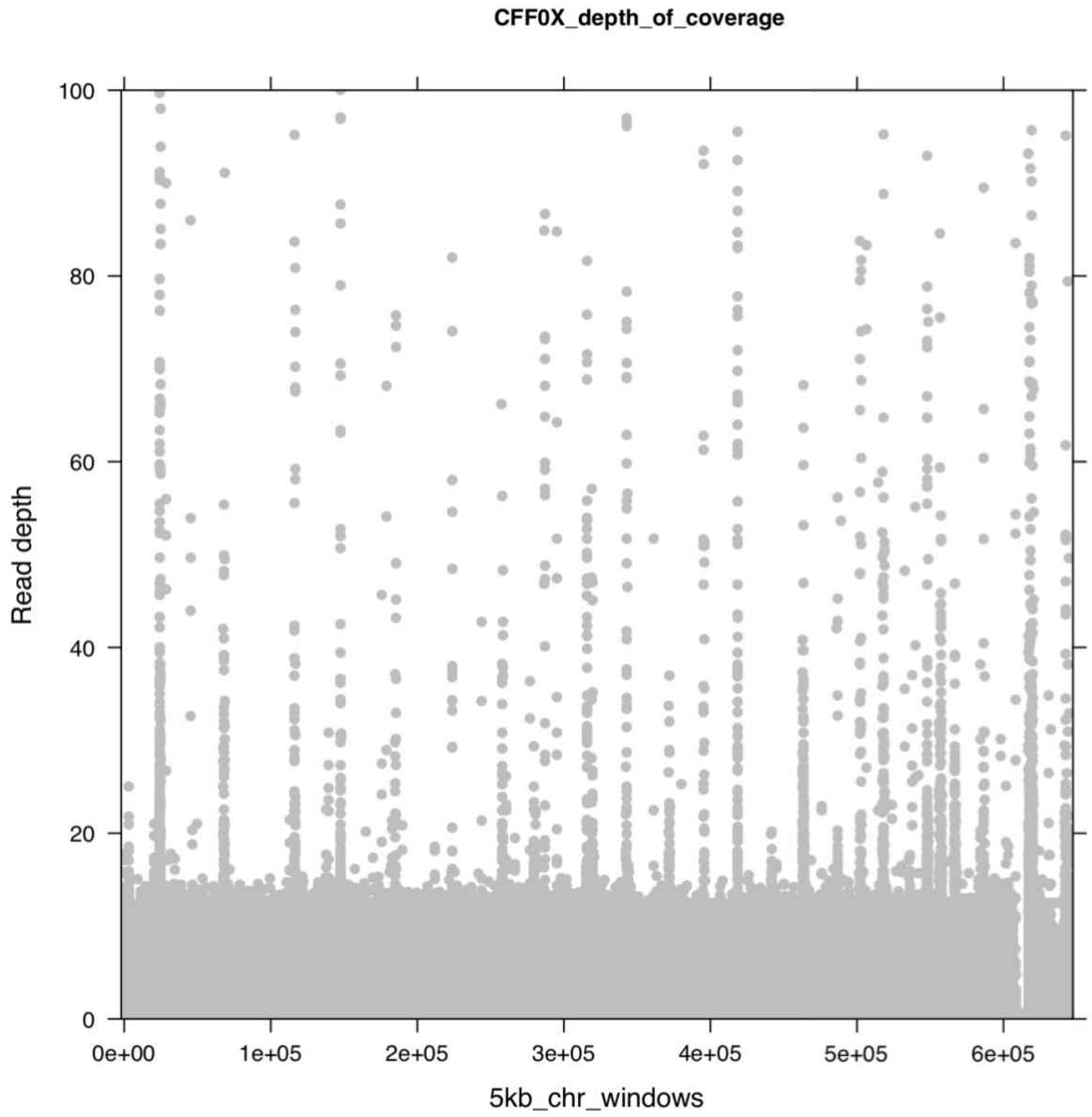


Figure B4: CFF0X depth of coverage.

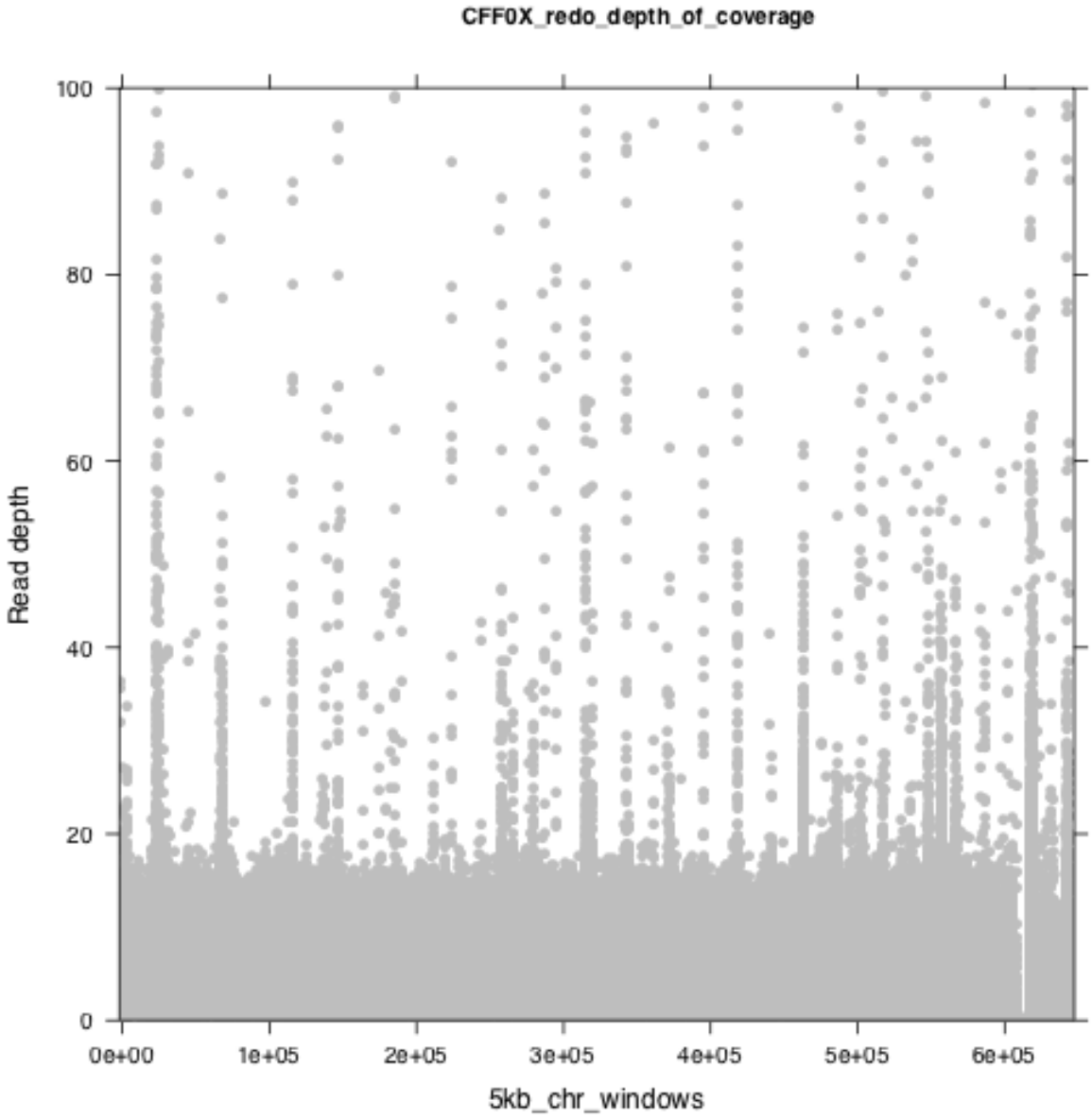


Figure B5: CFF0X (re-sequenced) depth of coverage.

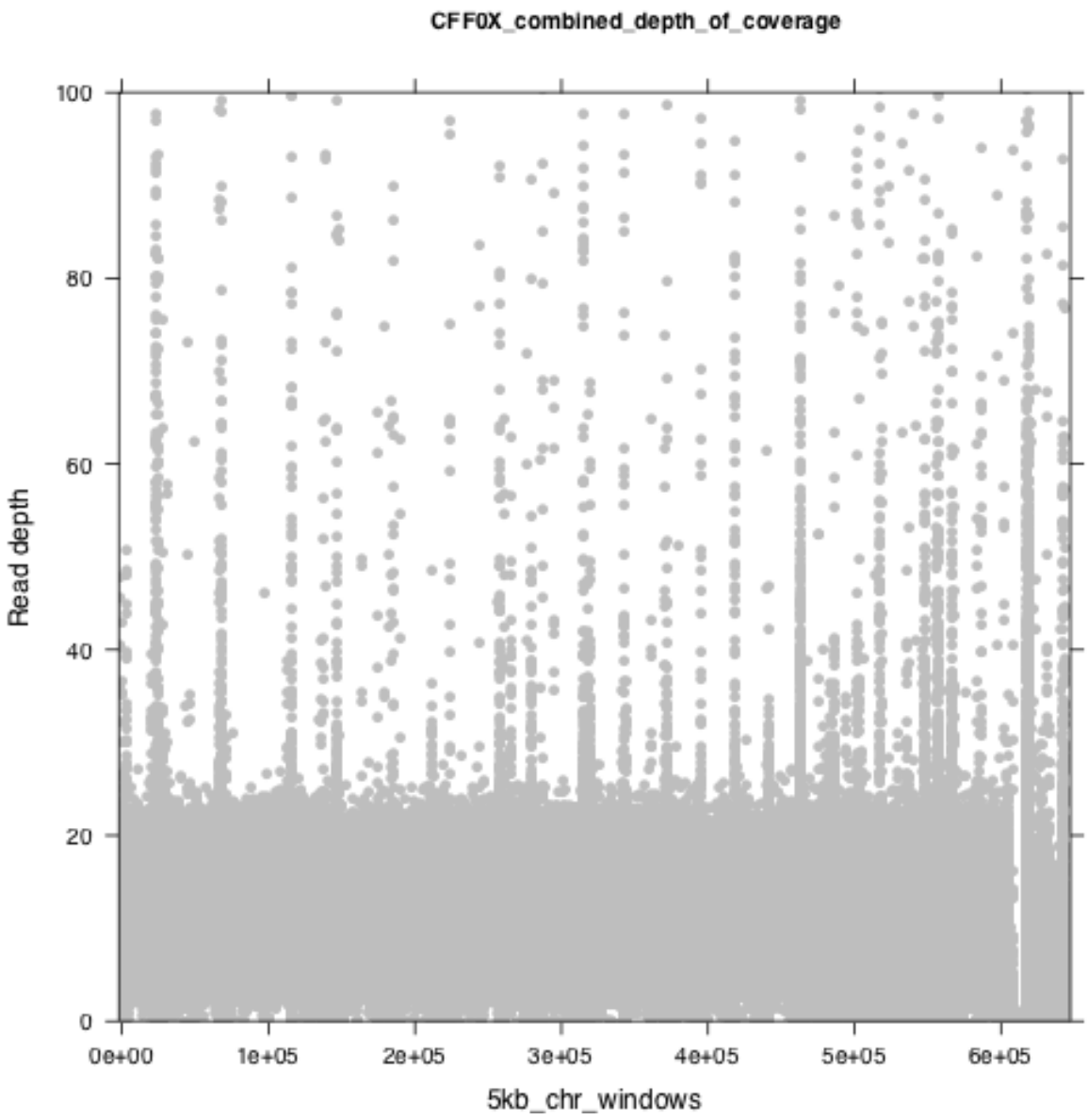


Figure B6: CFF0X (combined) depth of coverage.

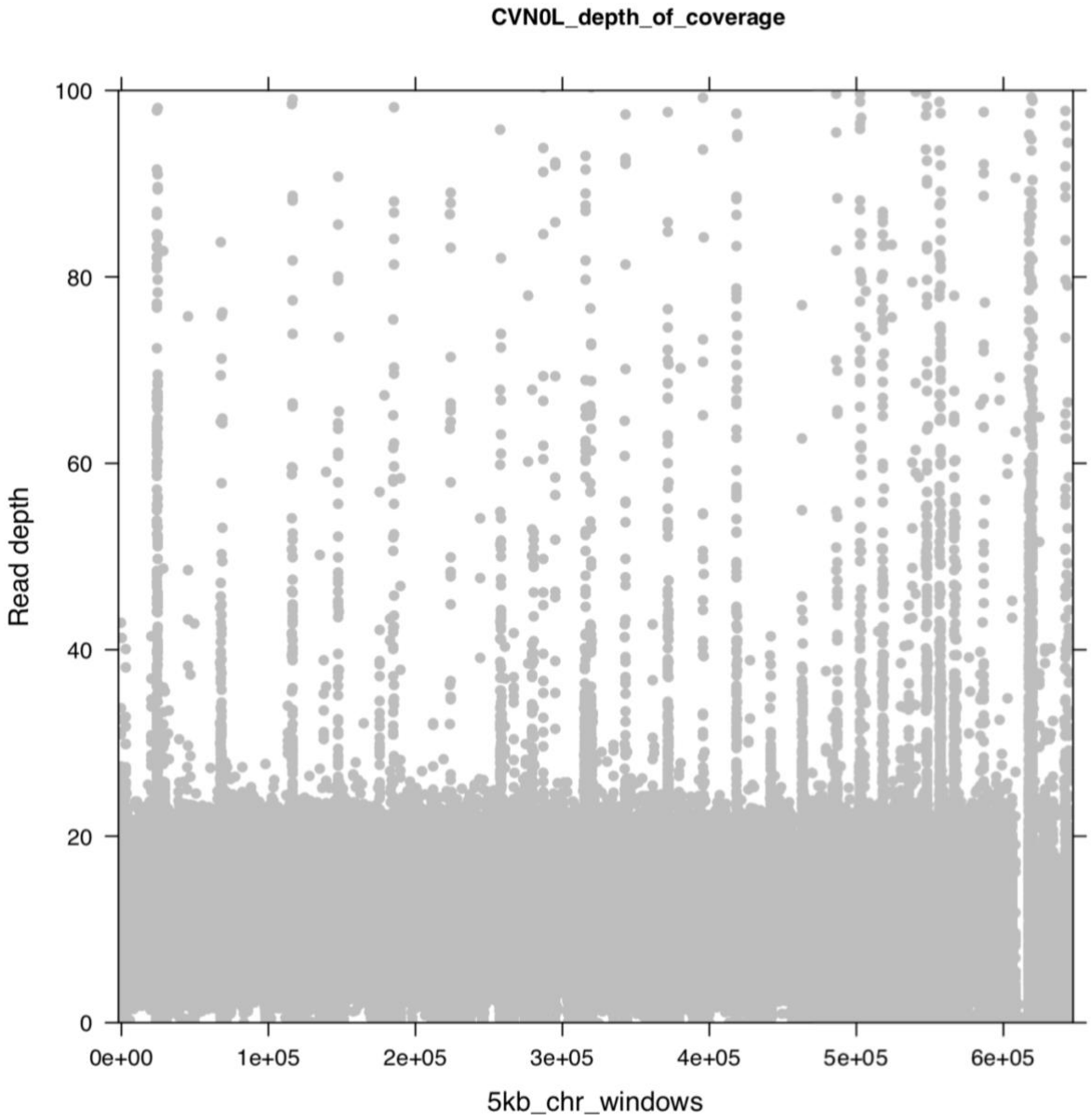


Figure B7: CVN0L depth of coverage.

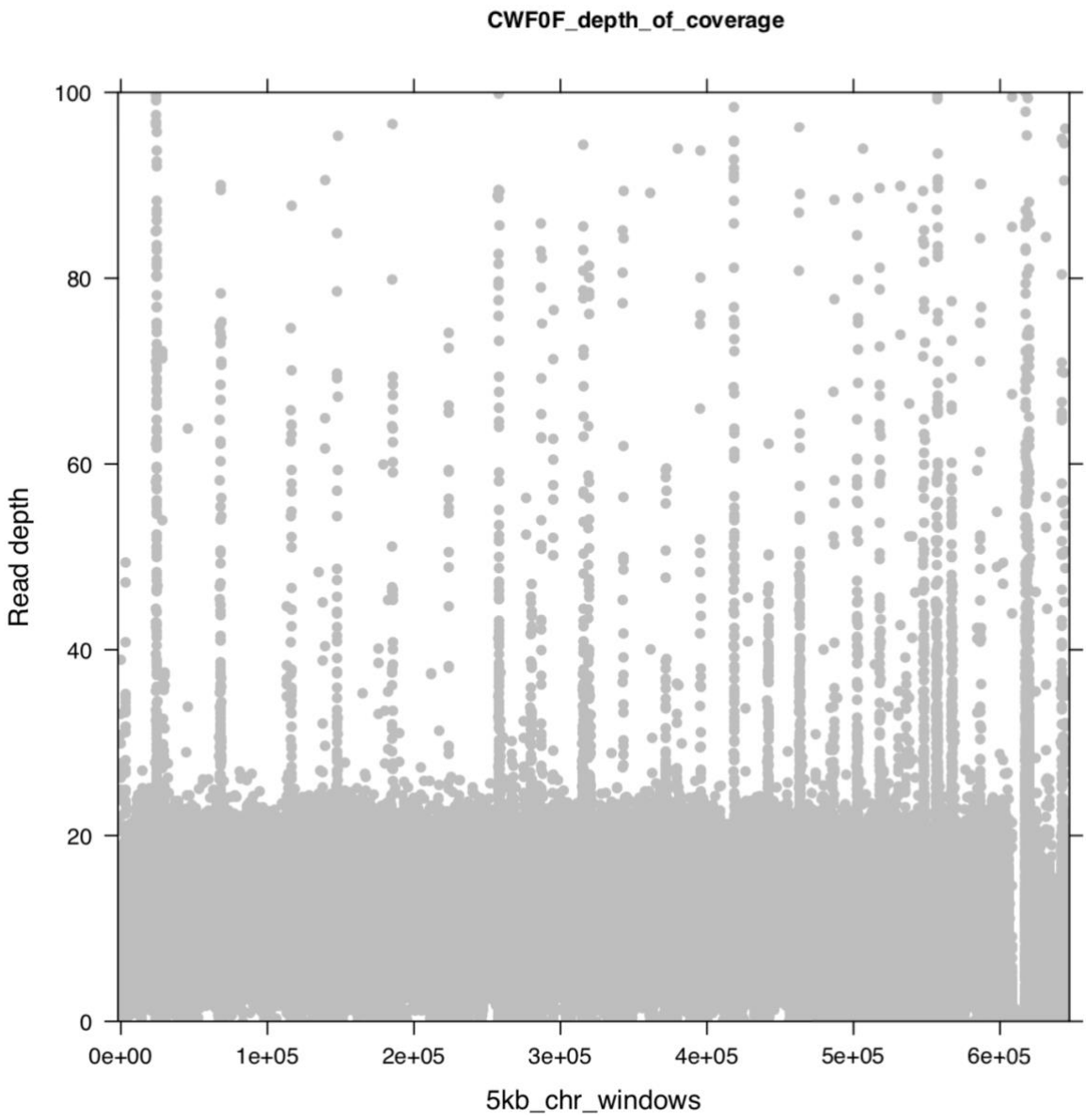


Figure B8: CWF0F depth of coverage.

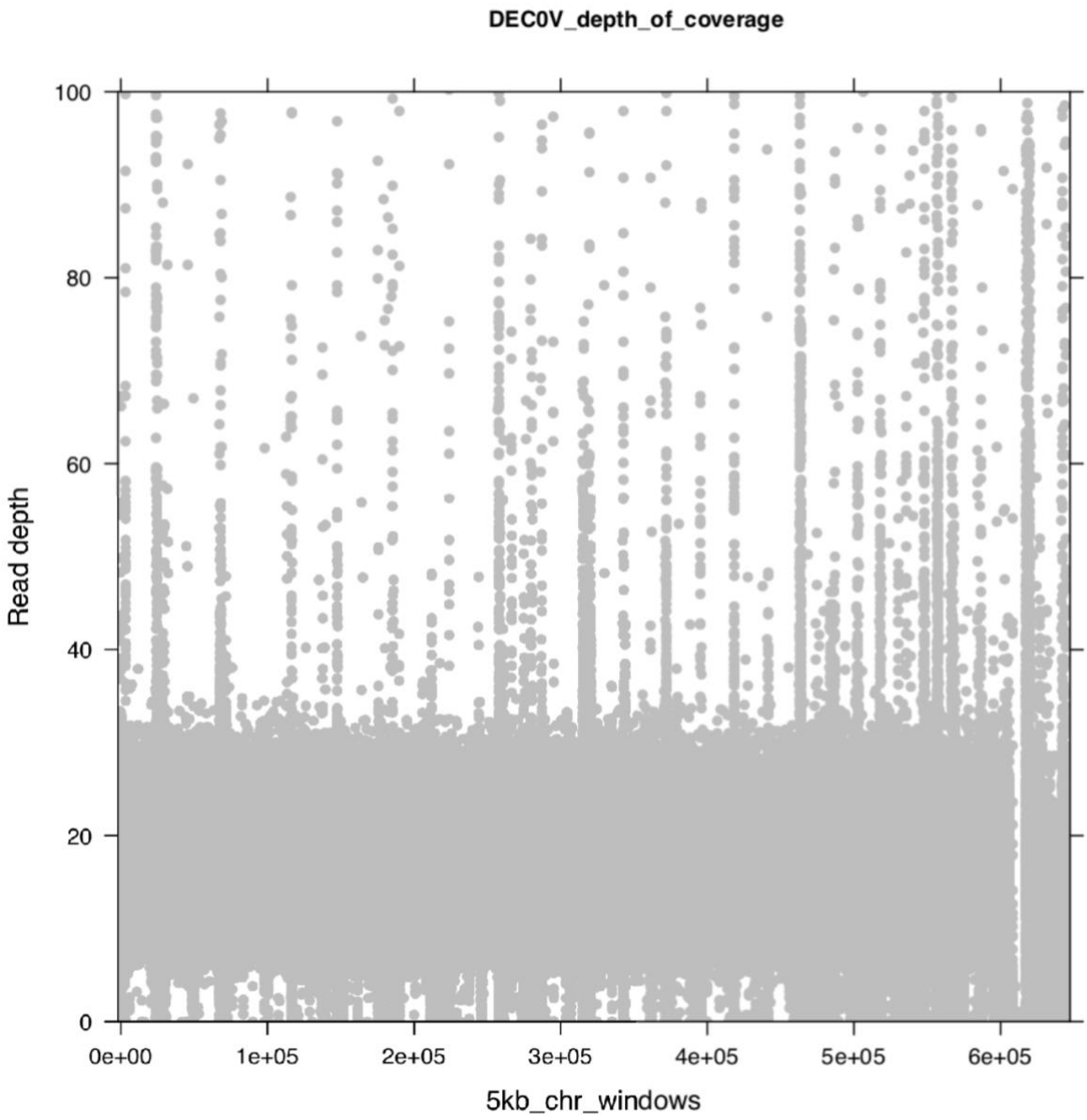


Figure B9: DEC0V depth of coverage.

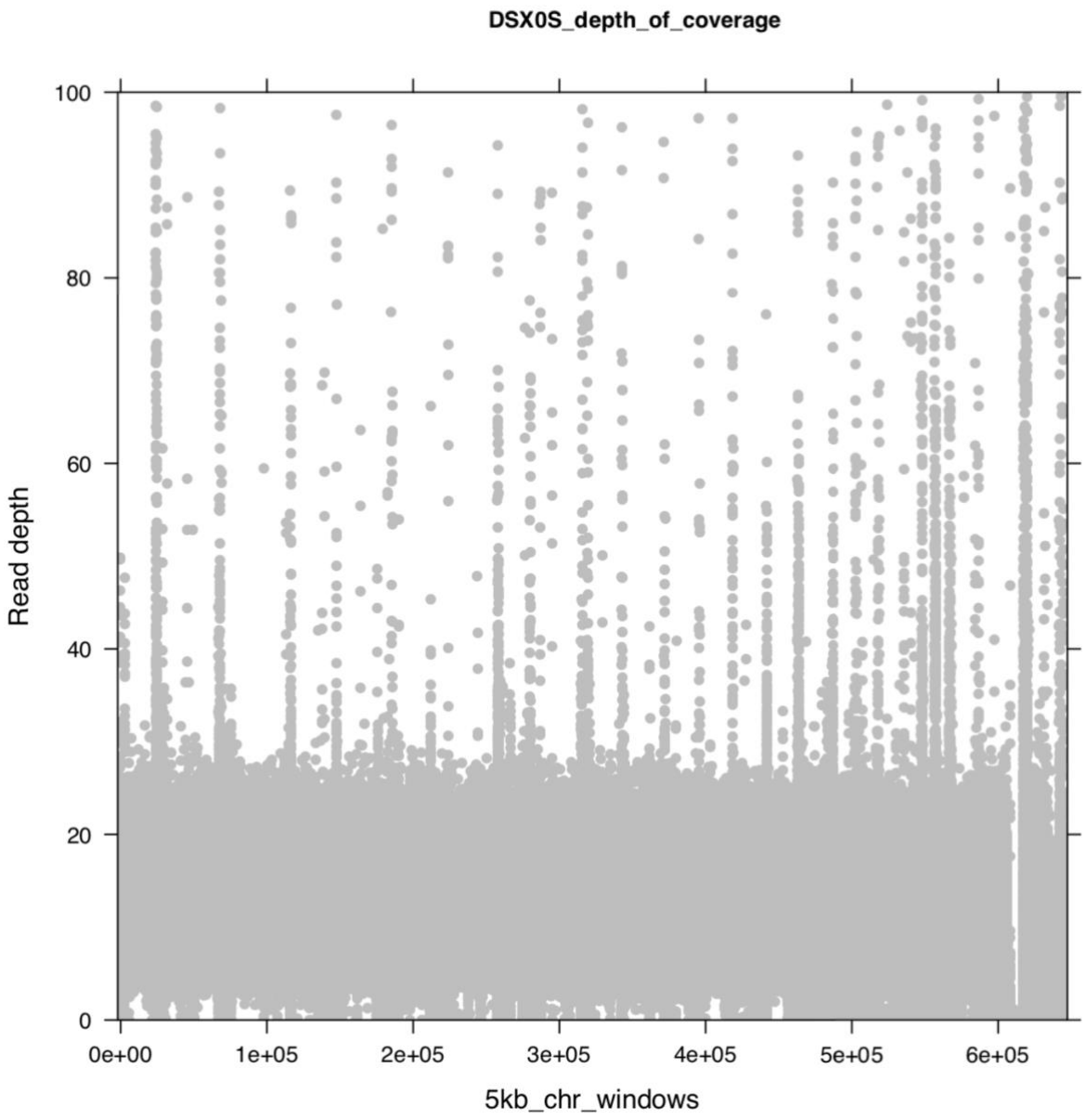


Figure B10: DSX0S depth of coverage.

APPENDIX C: Upset plots showing CNV intersections between samples of the same self-identified ethnolinguistic group.

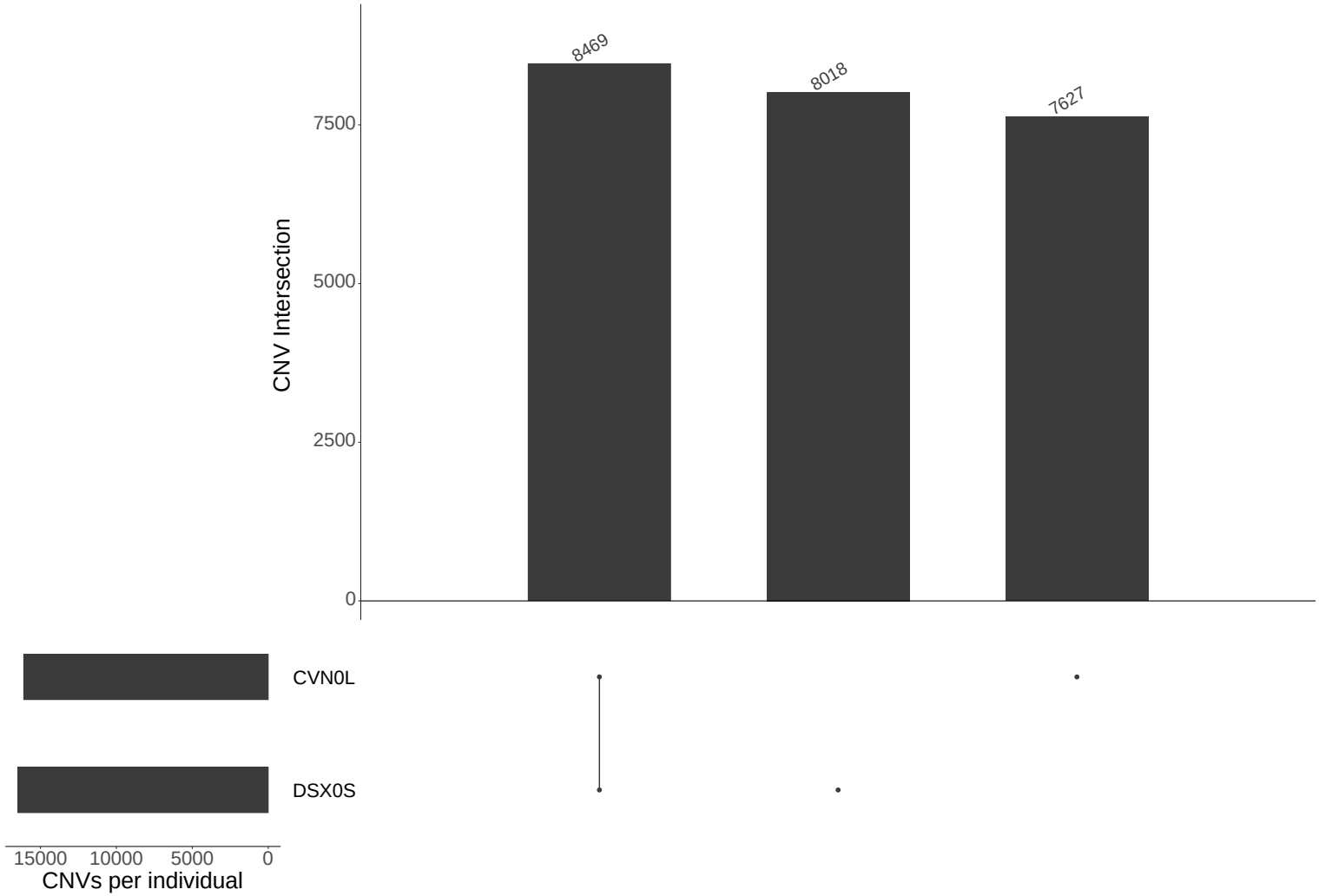


Figure C1: CNV overlap between CVN0L and DSX0S (Tswana samples).

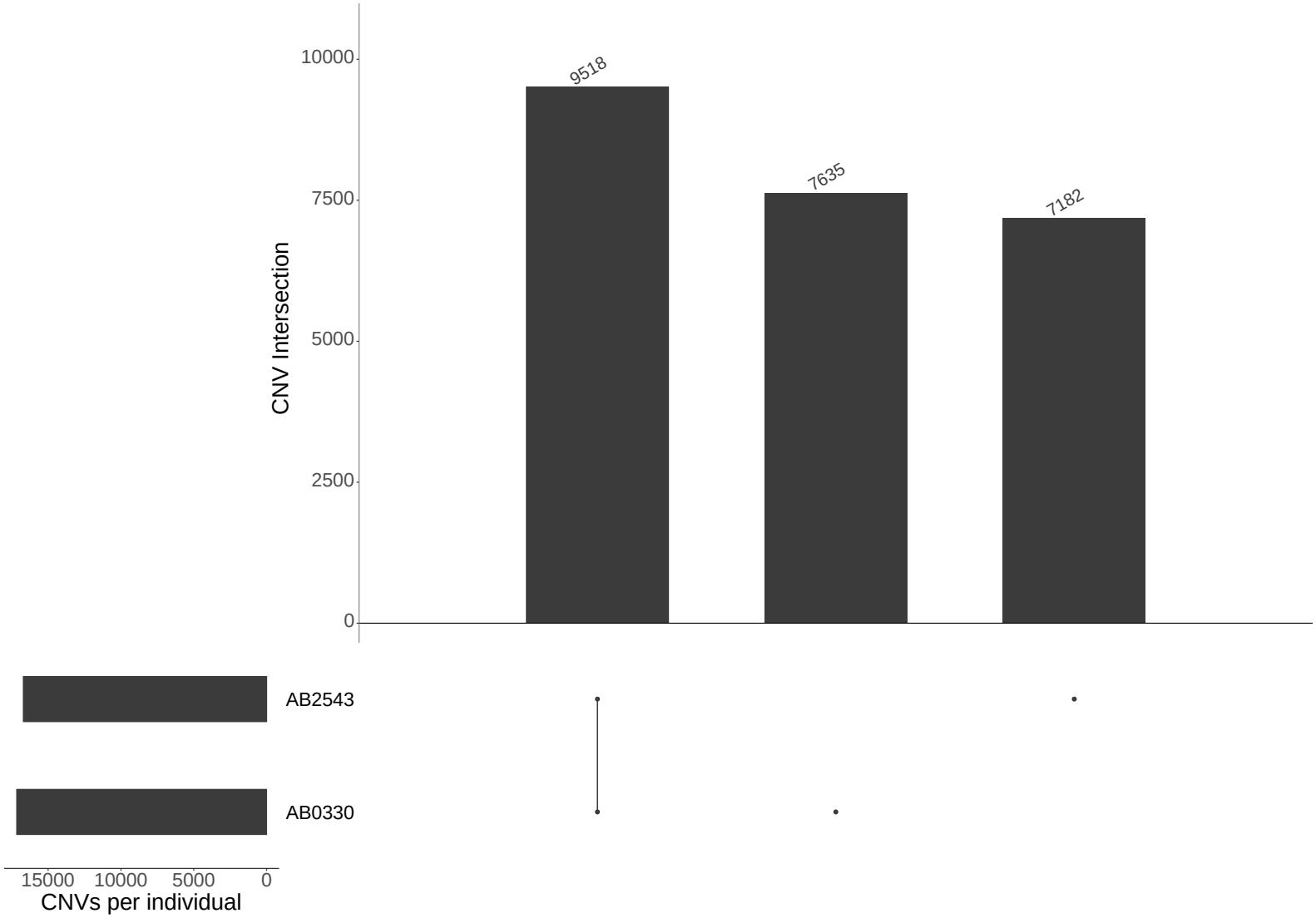


Figure C2: CNV overlap between AB2543 and AB0330 (Tsonga samples).

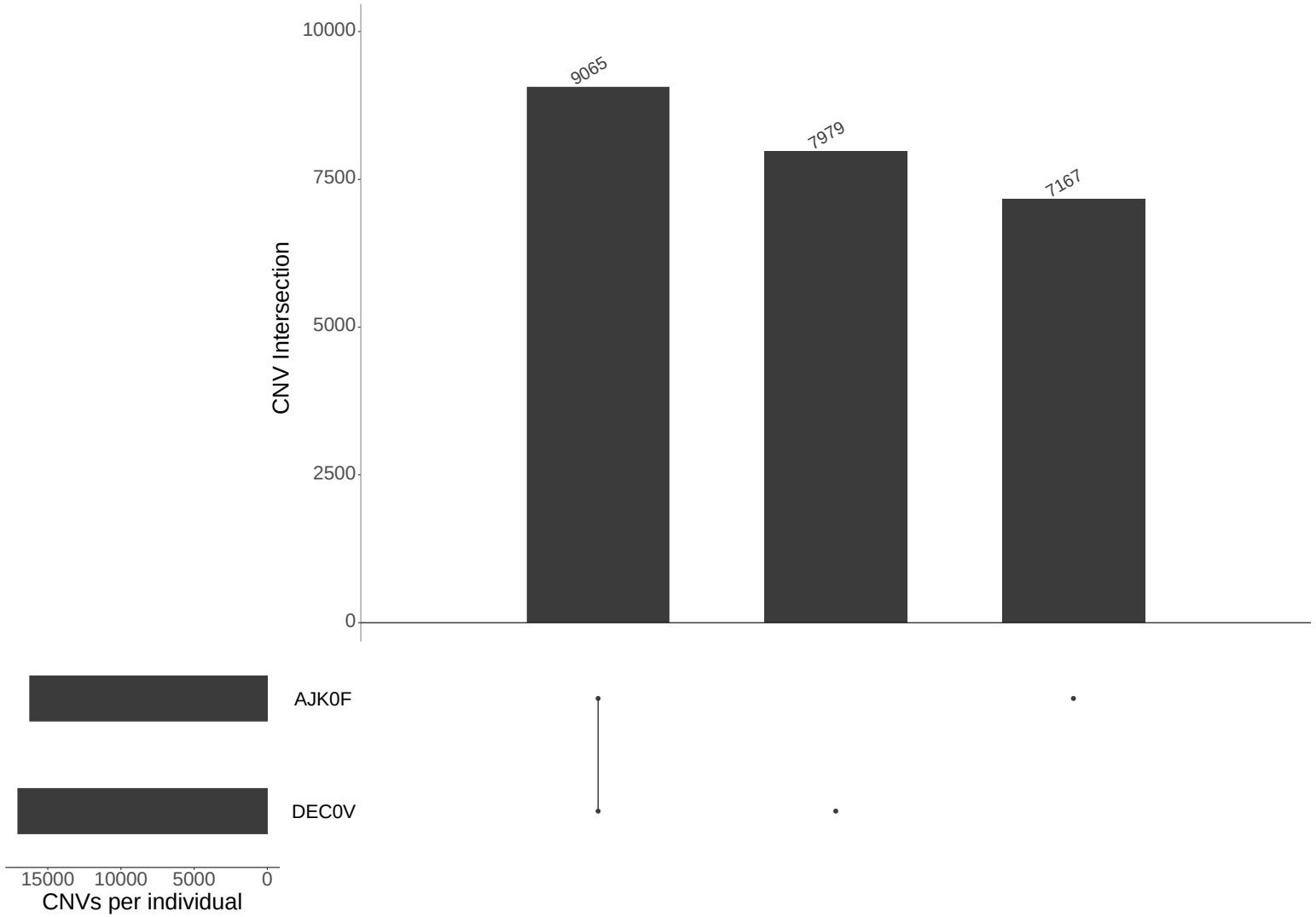


Figure C3: CNV overlap between AJK0F and DEC0V (Venda samples).

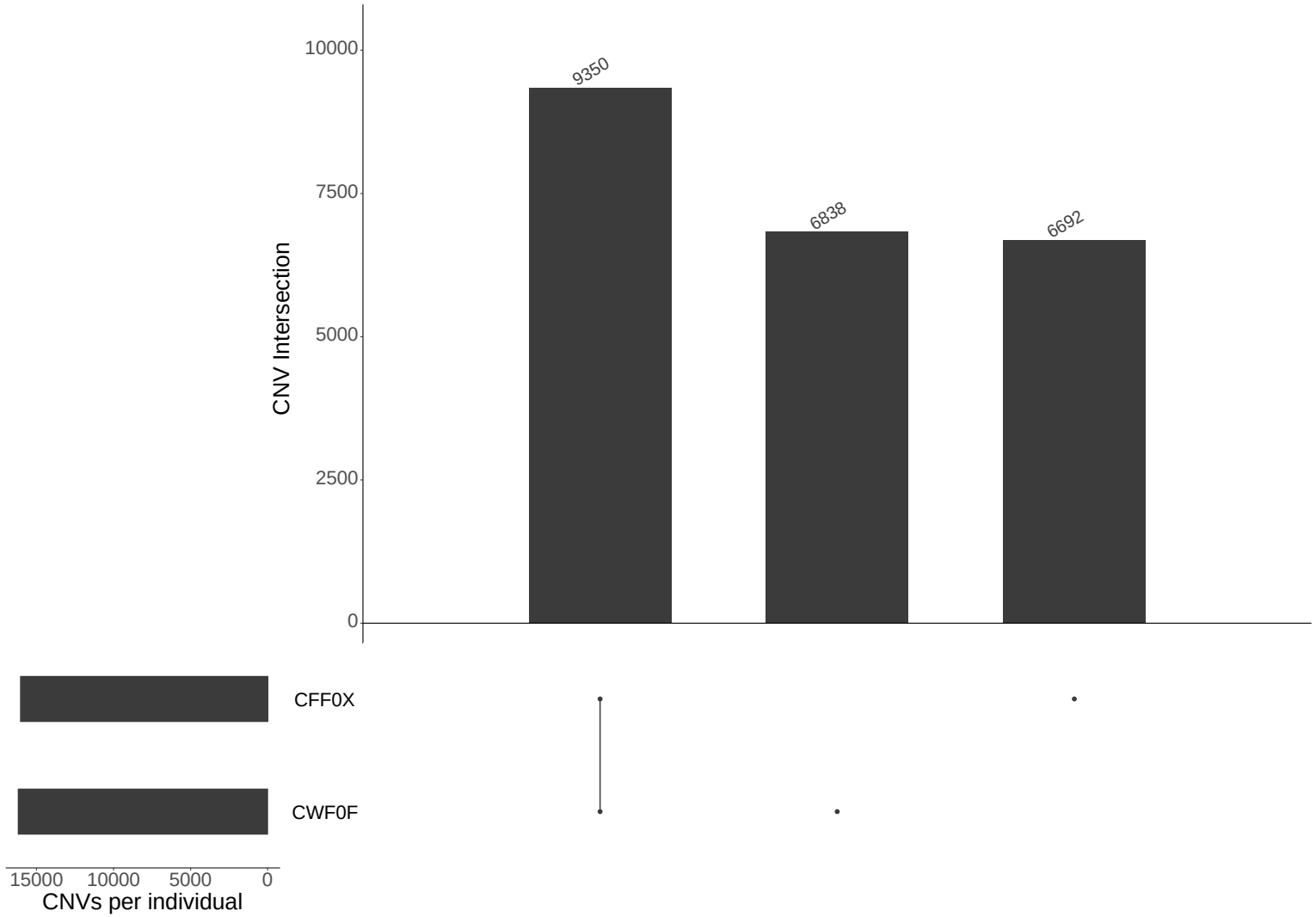


Figure C4: CNV overlap between CFF0X and CWF0F (Xhosa samples).

APPENDIX D: Approval of title.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Private Bag 3 Wits, 2050
Fax: 027117172119
Tel: 02711 7172076

Reference: Mrs Sandra Benn
E-mail: sandra.benn@wits.ac.za

03 January 2022
Person No: 810316
PAG

Ms LAA Cottino
P O Box 652161
Benmore
2010
South Africa

Dear Ms Laura Cottino

Doctor of Philosophy: Approval of Title

We have pleasure in advising that your proposal entitled *A systematic assessment of the Copy Number Variation (CNV) landscape in ADME genes in Sub-Saharan African populations* has been approved. Please note that any amendments to this title have to be endorsed by the Faculty's higher degrees committee and formally approved.

Yours sincerely

Mrs Sandra Benn
Faculty Registrar
Faculty of Health Sciences

APPENDIX E: Ethical clearance certificate (M180738).

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Miss Laura Cottino

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M180738

NAME: (Principal Investigator)	Miss Laura Cottino
DEPARTMENT:	School of Pathology Division of Human Genetics Sydney Brenner Institute for Molecular Bioscience (SBIMB) University of the Witwatersrand
PROJECT TITLE:	A systematic assessment of the Copy Number Variation (CNV) landscape in Absorption, Distribution Metabolism and Excretion (ADME) genes in sub-Saharan African populations
DATE CONSIDERED:	27/07/2018
DECISION:	Approved uncondtionally
CONDITIONS:	
SUPERVISOR:	Dr Venesa Sahibdeen and Prof Scott Hazelhurst
APPROVED BY:	CB Penny Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	07/12/2018

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **July** and will therefore be due in the month of **July** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date: 12 December 2018

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

APPENDIX F: Ethical clearance certificate for use of AWI-Gen data (array data and DNA for long-read sequencing) (classified as a sub-study under M170880) (M2009102).

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research & Post Graduate Affairs)

TO: Ms L Cottino
School of Pathology
Department of Human Genetics
Sydney Brenner Institute for Molecular Biosciences
Medical School
University

E-mail: lauracottino@gmail.com

CC: Supervisor: Professor S Hazelhurst <Scott.Hazelhurst@wits.ac.za>
and <HREC-Medical.ResearchOffice@wits.ac.za>

FROM: Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 2020/10/17

REF: *R14/49*

PROTOCOL NO: **M2009102** *(This is your ethics application study reference number. Please quote this reference number in all correspondence relating to this study)*

PROJECT TITLE: *A systematic assessment of the Copy Number Variation (CNV) landscape in ADME genes in sub-Saharan African populations*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to the Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Ms L Cottino

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M2009102

NAME: (Principal Investigator)	Ms L Cottino
DEPARTMENT:	School of Pathology Department of Human Genetics Sydney Brenner Institute for Molecular Biosciences Medical School University
PROJECT TITLE:	A systematic assessment of the Copy Number Variation (CNV) landscape in ADME genes in sub-Saharan African populations
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M170880
SUPERVISOR:	Professor S Hazelhurst
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2020/10/17

This clearance certificate is valid for 5 years from the date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the 3rd Floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **September** and will therefore reports and re-certification will be due early in the month of **September** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

16 October 2020
Date

APPENDIX G: Plagiarism declaration.



PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I LAURA ANN ALICE COTTINO (Student number: 810316) am a student registered for the degree of DOCTOR OF PHILOSOPHY in the academic year 2022.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: _____ Date: 22 February 2022

APPENDIX H: Acknowledgment of Turnitin report by supervisors.

SYDNEY BRENNER INSTITUTE FOR MOLECULAR BIOSCIENCE

School of Electrical and Information Engineering
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
Private Bag 3, 2050 Wits, South Africa
Tel: +27 11 717 6181
Fax: +27 86 553 6657
Email: scott.hazelhurst@wits.ac.za

14 February 2022

Postgraduate Studies Committee
Faculty of Health Sciences
University

Dear Professor Papathanasopoulos,

PhD submission – Laura Cottino – 810316 – Turnitin

We have considerd the Turnitin report for Ms Cottino's PhD and are satisfied with the originalitiry therein. According to the report there is a similarity of 20%. However, 7 of the 20 comes from a list of gene tables in the appendix which are necessary for the thesis. The rest of the similarity comes from very small amounts of similarity which match a variety of source material and is appropriately referenced.

Yours sincerely,

Scott Hazelhurst (Feb 14, 2022 15:14 GMT+2)

Scott Hazelhurst
Supervisor

Venesa Sahibdeen
Supervisor

APPENDIX I: Turnitin report.

Laura Cottino (810316) - PhD Thesis.docx

ORIGINALITY REPORT

20%	18%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	pharmaadme.org Internet Source	7%
2	www.nature.com Internet Source	1%
3	www.ncbi.nlm.nih.gov Internet Source	1%
4	www.frontiersin.org Internet Source	<1%
5	www.biorxiv.org Internet Source	<1%
6	academic.oup.com Internet Source	<1%
7	hdl.handle.net Internet Source	<1%
8	link.springer.com Internet Source	<1%
9	science.sciencemag.org Internet Source	<1%