Characterisation of pharmacogene allelic variation in African populations and development of a novel diplotype calling algorithm

David Twesigomwe

Supervisors:

Prof. Scott Hazelhurst

Prof. Zané Lombard

A Thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa, in fulfilment of the requirements of the degree of Doctor of Philosophy.

June 2022

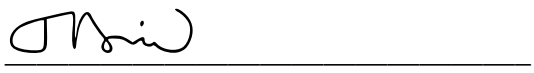
WITS UNIVERSITY

NATIONAL HEALTH LABORATORY SERVICE

SYDNEY BRENNER INSTITUTE FOR MOLECULAR BIOSCIENCE

Declaration

I, David Twesigomwe declare that this Thesis is my own, unaided work, unless otherwise specified in the text. 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 this or any other University. Co-author contributions to each published article or manuscript are listed under Author contributions.



Date: 6th June, 2022.

To my family, friends and mentors who have supported me on this journey,

and in memory of my mother.

Presentations arising from this thesis

1. Twesigomwe D., Lombard Z., and Hazelhurst S. on behalf of the Wits-H3A/GSK ADME Collaboration. Poster presentation: Characterisation of *CYP2D6* pharmacogenomic variation in African populations: an integrative bioinformatics approach. Pharmacogenomics Global Research Network–American Society of Human Genetics 2021 Symposium (PGRN-ASHG 2021). October 18, 2021.
2. Twesigomwe D., Drögemöller B.I., Wright G.E.B., Siddiqui A., da Rocha J., Lombard Z., and Hazelhurst S. Poster presentation: StellarPGx: A Nextflow pipeline for calling star alleles in cytochrome P450 genes. South African Society for Bioinformatics Student Coucil (SASBi-SC)/South African Genetics Society (SAGS) 2021 Virtual Student Symposium. September 14-16, 2021.
3. Twesigomwe D., Lombard Z., and Hazelhurst S. on behalf of the Wits-H3A/GSK ADME Collaboration. Flash talk: Characterisation of *CYP2D6* pharmacogenomic variation in African populations: an integrative bioinformatics approach. Eighteenth H3Africa Consortium Meeting – Hybrid meeting. August 30 – September 2, 2021.
4. Twesigomwe D., Drögemöller B.I., Wright G.E.B., Siddiqui A., da Rocha J., Lombard Z., and Hazelhurst S. Poster presentation: StellarPGx: A Nextflow pipeline for calling star alleles in cytochrome P450 genes. 29th Conference on Intelligent Systems for Molecular Biology/European Conference on Computational Biology (Track: Bioinformatics Open Source Conference - virtual). July 25-30, 2021.
5. Twesigomwe D., Drögemöller B.I., Wright G.E.B., Siddiqui A., da Rocha J., Lombard Z., and Hazelhurst S. Poster presentation: StellarPGx: A Nextflow pipeline for calling star alleles in cytochrome P450 genes. International Society for Computational Biology – African Society for Bioinformatics and Computational Biology (ISCB–Africa ASBCB) Conference on Bioinformatics (virtual). June 7-10, 2021.

6. Twesigomwe D., Drögemöller B.I., Wright G.E.B., Siddiqui A., da Rocha J., Lombard Z., and Hazelhurst S. Poster presentation: CypGen: A Nextflow pipeline for calling star alleles in cytochrome P450 genes. Biological Data Science Meeting (virtual), Cold Spring Harbor Laboratory. November 4–6, 2020.
7. Twesigomwe D., Lombard Z., and Hazelhurst S., on behalf of the Wits-H3Africa/GlaxoSmithKline ADME collaboration. Characterisation of *CYP2D6* pharmacogenomic variation in African populations: an integrative bioinformatics approach. Oral presentation: Faculty of Health Sciences (FHS) Research Day and Postgraduate Expo (virtual). October 15, 2020.
8. Twesigomwe D., Lombard Z., and Hazelhurst S., on behalf of the Wits-H3Africa/GlaxoSmithKline ADME collaboration. Characterisation of *CYP2D6* pharmacogenomic variation in African populations: an integrative bioinformatics approach. Poster presentation: 28th Conference on Intelligent Systems for Molecular Biology (virtual). July 13-16, 2020.
9. Twesigomwe D., Lombard Z., and Hazelhurst S., on behalf of the Wits-H3Africa/GlaxoSmithKline ADME collaboration. Characterisation of *CYP2D6* pharmacogenomic variation in African populations: an integrative bioinformatics approach. Oral presentation: International Society for Computational Biology (ISCB) Student Council Symposium (virtual). July 11-12, 2020.
10. Twesigomwe D., Lombard Z., and Hazelhurst S., on behalf of the Wits-H3Africa/GlaxoSmithKline ADME collaboration. Characterisation of *CYP2D6* pharmacogenomic variation in African populations: an integrative bioinformatics approach. Poster presentation: International Society for Computational Biology – African Society for Bioinformatics and Computational Biology (ISCB–Africa ASBCB) Conference on Bioinformatics. November 11-15, 2019.

11. Twesigomwe D., Lombard Z., and Hazelhurst S., on behalf of the Wits-H3Africa/GlaxoSmithKline ADME collaboration. Characterisation of *CYP2D6* pharmacogenomic variation in African populations: an integrative bioinformatics approach. Oral presentation: The 3rd ISCB–Africa Student Council Symposium. November 11, 2019.
12. Twesigomwe D., Lombard Z., and Hazelhurst S., on behalf of the Wits-H3Africa/GlaxoSmithKline ADME collaboration. Characterisation of *CYP2D6* pharmacogenomic variation in African populations: an integrative bioinformatics approach. Poster presentation: Southern African Society for Human Genetics Congress, Cape Town, South Africa. August 3–6, 2019.

Publications arising from this thesis

1. Twesigomwe D., Drögemöller B.I., Wright G.E.B., Siddiqui A., da Rocha J., Lombard Z. and Hazelhurst S. (2021). StellarPGx: A Nextflow pipeline for calling star alleles in cytochrome P450 genes. *Clinical Pharmacology and Therapeutics* 110(3), 741-749. doi: 10.1002/cpt.2173.
2. Twesigomwe D., Wright G.E.B., Drögemöller B.I., da Rocha J., Lombard Z. and Hazelhurst S. (2020). A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: A focus on *CYP2D6* genotyping. *npj Genomic Medicine* 5, 30. doi.org/10.1038/s41525-020-0135-2.

Manuscripts in preparation

3. Characterisation of *CYP2D6* pharmacogenomic variation in African populations: An integrative bioinformatics approach.
4. Characterisation of the *CYP2B6* and *CYP2A6* allelic variation in African populations.

Abstract

Genetic variation is in part responsible for the variability in drug response within and between populations. Therefore, there is a shift in drug therapy in the form of implementing genotype-guided treatment in place of the 'one-size-fits-all' model so as to promote drug efficacy and reduce the risk of adverse drug reactions. However, the full catalogue of pharmacogenetic variants is yet to be established, in particular for African populations. Therefore, this study aimed to characterise the variation in three core pharmacogenes in drug metabolism i.e. *CYP2D6*, *CYP2B6* and *CYP2A6*, across diverse continental African populations.

Given the known challenges of genotyping these hypervariable genes, a benchmark of the state-of-the-art star allele calling bioinformatics algorithms was performed using simulated and real-world whole genome sequence (WGS) data. Furthermore, we developed a novel graph- and Nextflow-based pipeline (StellarPGx) to facilitate scalable, reproducible and more accurate CYP genotyping using WGS data. Thereafter, we applied a consensus star allele calling approach, involving StellarPGx and the other existing tools, to assess the *CYP2D6*, *CYP2B6* and *CYP2A6* allelic diversity mainly across sub-Saharan Africa, based on 962 high-depth African genomes. Targeted Single Molecule Real-Time (SMRT) Sequencing was used to validate novel *CYP2D6*, *CYP2B6* and *CYP2A6* haplotypes found in individuals whose DNA samples were available to us through various collaborations.

Our analysis highlighted the frequency of known normal, decreased, increased and no-function *CYP2D6*, *CYP2B6* and *CYP2A6* alleles across sub-Saharan Africa in comparison with other global populations. Differences in predicted CYP2D6 and CYP2B6 metaboliser phenotypes were observed even between some populations from neighbouring geographical regions. In addition, from the short-read WGS data from African individuals, we inferred over 77 potential novel African-specific alleles in *CYP2D6*, *CYP2B6* and *CYP2A6* combined. Targeted SMRT sequencing for a subset of the study population enabled further characterisation of two novel African-specific alleles apiece for *CYP2D6*, *CYP2B6* and *CYP2A6*.

This study highlights the importance of combining multiple tools in accurately mining high coverage WGS data for variation in highly polymorphic CYP genes. Furthermore, our novel bioinformatics tool, StellarPGx, addresses the need to implement a pipeline for accurate, scalable, and reproducible pharmacogene allele calling using WGS data. The variability in pharmacogene allele and phenotype distributions across the African populations represented in this study, and the identification of several novel alleles underscore the need for investigating pharmacogene variation in an African context to reliably inform clinical pharmacogenomics implementation strategies across Africa.

Acknowledgements

I would like to express my heartfelt gratitude and appreciation to the following people and institutions that have made this PhD possible:

To my supervisors, Prof Zané Lombard and Prof Scott Hazelhurst, for their invaluable support and advice during my research, and also for their career mentorship. Thank you for sharing your expertise and wisdom with me. You are wonderful role models!

To my family for their unending love, care, support and guidance throughout my studying years. Even though I was based far away from home during this PhD journey, I felt connected to you every step of the way. Special gratitude to my parents Rose and D. K. Byamugisha for their unconditional love and motivation. To my biological mother, Donam, thank you for bringing me into this world and for fighting for me to live early on in my life — may you rest in peace dear mum. To my wonderful siblings, Rachael, Franklin, Liz, Helen, Daniel, Ambrose, and Denis, thank you for always caring for me, sharing words of advice and encouragement, and also inspiring me with your own successes. A big special shout out to Gilbert, Sylvia, Aunt Joan, Uncle Tom, Aunt Oliver, Uncle Geofrey, Sharon, Sam, Beckie, Ruth, Dora, Diana, Ananda, Daniella, Divine, Lilian, Vincent, Uncle Jeff, Aunt Immaculate, and to everyone in our extended family — I am so blessed to have you in my life.

To my teammates on the Wits-H3A/GSK ADME project, I am incredibly lucky to have met and shared so much with you. Jorge, thanks for your amazing selflessness and for the countless hours we spent bouncing ideas off of each other. Houcem, Laura, Blessing, Nhlamulo, Samah, Chaimae, Gerrit, Sumir, Mamana, and to every other awesome colleague on the project, thank you for teaching me so much in such a short period of time and for sharing your wisdom with me. Special shout out to Britt and Galen for their expert advice throughout my PhD, contributions to all manuscripts, and for always taking time from their busy schedules to help me improve as a young scientist in pharmacogenomics.

To the incredible Sydney Brenner Institute for Molecular Bioscience (SBIMB) family, thank you for embracing me from the very start and for the support throughout my PhD research, and development as a scientist. Special gratitude goes to:

- Prof Michele Ramsay for creating an environment at the SBIMB that would be a dream for every postgraduate student across the globe, and in particular for her great feedback and advice pertinent to my PhD research.
- Phelelani and Shaun for always taking time out of their busy schedule to help out with any bioinformatics problems, and above all for your kindness and true friendship.
- The SBIMB Biobank team (Cassandra, Busisiwe, Natalie) without whom my laboratory

validation work would have been an insurmountable mountain.

- The SBIMB projects office for their awesome support that enabled me to focus fully on my PhD research — Jocelyne, Mmatshepo, Yusuf and Thabo, I wish many blessings upon you for always being there for all the students at the SBIMB.
- Prof Collen Masimirembwa for the expert advice in pharmacogenomics and for being such a great role model.
- Mahtaab, Jorge, Surina, Laura, Micaela, Evans, Hleli, Runesu, Michelle, Carl, Micaella, Heather, Blessing, Voke, Shingie, Theo, Ananyo, Dhriti, Henry (and literally every student/ postdoc/staff at the SBIMB) for the friendship and camaraderie that helped me settle in as an international student and also for your invaluable advice pertinent to my bioinformatics training and PhD research.

To the Division of Human Genetics, NHLS, Wits, for the incredible support you gave me at the beginning of my studentship and for the entirety of my PhD research.

To Dr Stuart Ali for sharing his time, experience and advice with me in the early parts of my laboratory validation assays.

To Dr Andrea Gaedigk for your expert pharmacogenomics advice and guidance regarding genotyping *CYP2D6* and other complex pharmacogenes.

To all my mentors — formal and informal, thank you so much for believing in me and encouraging me to follow my passion for bioinformatics. Special thanks to the following:

- Dr John Kitayimbwa for changing my life by introducing me to bioinformatics and computational biology after agreeing to meet me, without an appointment, and hearing out my story.
- Dr Jonathan Kayondo, Dr Daudi Jjingo, and Prof Alison Elliott for supporting me during my strategic internship at the Uganda Virus Research Institute which contributed to my bioinformatics training and soft skills that came in handy in this PhD research.
- Dr Gerald Mboowa for being the best mentor ever in bioinformatics and research in general, and for your rapid responses whenever I contacted you with any challenges.

To Westridge Study Centre for supporting me for the best part of two years during my PhD and helped me deal with the cultural shocks around Johannesburg. Father Manu, Thiago and John Paul, thank you for helping me with my spiritual formation during this time. Massive shout out to all Westridge residents, alumni and friends for being there for me during this time.

To my friends Yuval, Nikkie, the Kharvas, Haroon, Hammaad, and Liesl — even though we met mostly for tennis, your keen interest in my research always gave me fresh perspectives and renewed energy.

To all my friends not mentioned above, thank you for caring about me and supporting me. Stay blessed.

To the GlaxoSmithKline (GSK) Research and Development Ltd for funding my PhD research under the Wits-H3A/GSK ADME collaboration. GSK was not responsible for the study design, and the views and opinions expressed in the publications arising from this thesis are not necessarily those of GSK.

To the National Research Foundation (NRF) of South Africa for awarding me a free standing bursary. The financial assistance of the NRF (Grant number: 128895) that supported me throughout this research is hereby acknowledged. The opinions, findings and conclusions or recommendations expressed in any publication generated by this NRF-supported research are those of the authors, and are not necessarily to be attributed to the NRF.

Lastly (but definitely not least!), many thanks to the study participants for your invaluable contribution in making this PhD research possible.

Table of Contents

List of Abbreviations

ABC	ATP binding cassette
ACB	African Caribbean in Barbados
aCGH	Comparative Genomic Hybridisation arrays
ADME	Absorption, Distribution, Metabolism, Excretion
ADRs	Adverse Drug Reactions
AF	Allele Frequency
AIDS	Acquired Immune Deficiency Syndrome
AMR	Admixed American
AS	Activity Score
ASW	African Ancestry in South West USA
AWI-Gen	Africa Wits-INDEPTH partnership for Genomics studies
BAM	Binary Alignment Map
BFA	study participants in Burkina Faso
BOT	Batswana participants
BRN	Berom in Nigeria
BSZ	Bantu-speakers in Zambia
bp	base pairs
CAM	study participants from Cameroon
CCS	Circular Consensus Sequence
CDC	Centers for Disease Control and Prevention
cDNA	complementary DNA
CFTR	Cystic Fibrosis Transmembrane conductance Regulator
CLIA	Clinical Laboratory Improvement Amendments
CN	Copy Number
CNV	Copy Number Variation
CPIC	Clinical Pharmacogenetics Implementation Consortium

CYP	Cytochrome P450
DALYs	Disability-Adjusted Life Years
dbSNP	Single Nucleotide Polymorphism Database
ddNTPs	dideoxynucleotide triphosphates
ddPCR	digital droplet PCR
DILI	Drug-Induced Liver Injury
DPWG	Dutch Pharmacogenetics Working Group
EAS	East Asian
EGAD	European Genome-Phenome Archive Dataset
ESN	Esan in Nigeria
EUR	European
ex	exon
FDA	Food and Drug Administration
FMO	Flavin-containing Monooxygenase
FNB	Fon in Benin
Frag	Fragment
HAAD	High coverage African ADME Dataset
H3A	Human Heredity and Health in Africa
GBD	Global Burden of Disease
GDF	GATK Depth-of-coverage Files
GeT-RM	Genetic Testing Reference Materials Coordination Program
GHA	study participants in Ghana
GRCh37	Genome Reference Consortium Human Build 37
GRCh38	Genome Reference Consortium Human Build 38
GVCF	Genomic Variant Call Format
GWAS	Genome Wide Association Studies
GWD	Gambian in Western Division — Mandinka
Hi-C	High-throughput chromosome conformation Capture

HiFi	High Fidelity
HIV	Human Immunodeficiency Virus
HTS	High-Throughput Sequencing
IM	Intermediate Metaboliser
Indels	Insertions and Deletions
int	intron
LD	Linkage Disequilibrium
LWK	Luhya in Webuye Kenya
mRNA	messenger RNA
miRNA	microRNA
MPA	Multiplex Probe Amplification
MSL	Mende in Sierra Leone
N	Number of samples
NCDs	Non-Communicable Diseases
NGS	Next Generation Sequencing
NM	Normal Metaboliser
NMR	Nicotine Metabolic Ratio
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
PD	Pharmacodynamics
PharmVar	Pharmacogene Variation Consortium
PharmGKB	Pharmacogenomics Knowledge Base
PGx	Pharmacogenomics
PK	Pharmacokinetics
PM	Poor Metaboliser
QC	Quality Control
SAHGP	Southern African Human Genome Programme

SAS	South Asian
SBS	Sequencing by Synthesis
SEB	south-eastern Bantu
SGDP	Simon's Genome Diversity Project
SLCO	Solute Carrier Organic anion transporter
SMRT	Single Molecule Real-Time
SNPs	Single Nucleotide Polymorphisms
SNVs	Small Nucleotide Variations
SSA	sub-Saharan Africa
SVs	Structural Variants
UGT	UDP-glucuronosyltransferase
UM	Ultrarapid Metaboliser
UTR	Untranslated Region
VCF	Variant Call Format
VIP	Very Important Pharmacogene
VKORC	Vitamin K epOxide Reductase Complex
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing/Sequence
WHO	World Health Organisation
Wits	University of the Witwatersrand
XL-PCR	Long-Range Polymerase Chain Reaction
YLLs	Years of Life Lost
YLDs	Years Lived with Disability
YRI	Yoruba in Nigeria

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Rationale

Modern humans originated in Africa as far back as >200,000 years ago, and their dispersal out of Africa is known to have occurred within the past 100,000 years (Nielsen et al., 2017; Tishkoff et al., 2009). In keeping with the out-of-Africa model, African populations have the highest levels of genetic diversity compared to any other present day continental population (Gomez et al., 2014; Tishkoff et al., 2009). The genetic diversity in African populations has been shaped by a number of factors mainly including migrations, admixture, gene flow, response to human disease, dietary differences, and other environmental features such as geography-based isolation (Choudhury et al., 2020; Gurdasani et al., 2015; Campbell and Tishkoff, 2008).

Over the past 25 years, the evolution of DNA sequencing – determining the order of DNA's nitrogenous building blocks i.e. Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) – has enabled scientists to study variant distributions across various populations (Heather and Chain, 2016; Auton et al., 2015; Sudmant et al., 2015). However, only a small fraction of the ~2000 ethnolinguistic groups on the African continent have been genetically characterised, which is in stark contrast to the numerous studies on populations of European and Asian ancestry (A. R. Bentley et al., 2020; Martin et al., 2019).

There are four major language families in Africa: Niger-Congo (includes the Bantu language family – spoken across a broad region of Africa), Afro-Asiatic (spoken predominantly in Saharan, north-eastern, and some parts of eastern Africa), Nilo-Saharan (spoken predominantly in Sudanic, Saharan, and eastern Africa), and the Khoe and San languages (Tishkoff et al., 2009). In addition to being genetically distant to non-African populations, genetic differences between these African ethnolinguistic groups are also quite profound (Choudhury et al., 2020; Tishkoff et al., 2009).

Investigating the patterns of genetic diversity in Africa is critical to not only understanding the peopling of African populations but also characterising the variability in human disease susceptibility, immunity, and response to medications among other aspects (A. R. Bentley et al., 2020; Choudhury et al., 2020; Kamp et al., 2021). In particular, addressing the underrepresentation of African populations in genomics research is pivotal to the advancement of precision medicine — an emerging approach for disease diagnosis, treatment and prevention that takes into account individual genetic variability alongside environmental and lifestyle factors (Caudle et al., 2016).

The effective implementation of precision medicine approaches across Africa necessitates extensive prior pharmacogenetics and pharmacogenomics research efforts aimed at investigating the relationship between genetic variation in multiple African ethnolinguistic groups and the response to various medications. Numerous associations between genetic variations and significantly decreased drug efficacy or increased risk of adverse drug reactions have been established based predominantly on studies involving individuals of European and Asian ancestry (PharmGKB gene-specific information tables, 2021). However, equivalent studies on African populations are few and far between (Rajman et al., 2017), which is quite staggering given the vast genetic diversity in Africa.

The current study forms part of a larger project focused on characterisation and functional analysis of variations present in the ADME genes in African populations (da Rocha, Othman, Botha, et al., 2021; Othman et al., 2021; da Rocha, Othman, Tiemessen, et al., 2021). ADME genes encode key proteins involved in the absorption, distribution, metabolism and excretion of drugs. The motivation for the larger ADME study as a whole is the need to expand the nascent pharmacogenomics research capacity in the Human Heredity and Health in Africa (H3Africa) Consortium (`https://h3africa.org`) through mining the whole genome sequence (WGS) data generated by H3A projects (see subsubsection 2.3.3) for functionally relevant variants or haplotypes in ADME genes. Another key component of the larger ADME project is the construction of portable and robust bioinformatics pipelines that can be applied on similar datasets by other pharmacogenomics researchers in Africa and beyond. This thesis considerably contributes to addressing these two overarching aspects.

The contribution of this thesis is threefold:

1. Benchmarking state-of-the-art algorithms for genotyping important ADME genes using high-throughput sequencing (HTS) data.
2. Addressing key deficiencies associated with existing algorithms by the development of a novel diplotype calling algorithm and subsequent application of a consensus allele assignment approach leveraging complementary tools to analyse novel African HTS data. A diplotype represents the two haplotypes (one on each chromosome) for a particular gene in an individual's genome.
3. Characterisation of the allelic variation in highly polymorphic and structurally variant pharmacogenes (*CYP2D6*, *CYP2B6* and *CYP2A6*) in African populations using HTS data, followed by laboratory validation of novel pharmacogenomically relevant alleles.

In particular, this thesis seeks to investigate the variation present in the *CYP2D6*, *CYP2B6* and *CYP2A6* genes in diversely sampled African populations using WGS data. The CYP2D6, CYP2B6 and CYP2A6 enzymes are important for the metabolism of several drugs (see `https://`

www.pharmgkb.org) including antidepressants, antiretrovirals, anticancer agents, nicotine, antimalarials and opioids to mention but a few. Genetic variation in all the three genes, some of which is known to affect the response to some of the essential medications (e.g. efavirenz and tamoxifen), has been extensively studied (Nofziger et al., 2020; Desta et al., 2021; Wassenaar et al., 2016). However, due to the limitations in the number of studies and sample sizes, critical gaps remain as regards to characterising the *CYP2D6*, *CYP2B6* and *CYP2A6* allelic variation in African populations, thus limiting the implementation of relevant genotype-guided therapy in Africa. Of note, *CYP2D6*, *CYP2B6* and *CYP2A6* are located in genomic loci that are difficult to interrogate using HTS approaches (Nofziger and Paulmichl, 2018; Desta et al., 2021; Wassenaar et al., 2016) and have often been considered to be part of the "inaccessible genome" (Drögemöller, Wright, et al., 2013; Auton et al., 2015).

The motivation for this thesis was, therefore, addressing the paucity of information on African-specific variation in these key pharmacogenes and the need to implement an accurate, scalable, and reproducible HTS-based allele calling bioinformatics pipeline.

The haplotypes (star alleles) — clusters of variants inherited together — in these complex pharmacogenes mainly comprise single nuclotide polymorphisms (SNPs), which are alterations at single base positions in the DNA sequence among individuals, as well as small insertions and deletions (indels) that could involve up to 50 DNA base pairs (Mills et al., 2006). SNPs and indels are also collectively referred to as small nucleotide variations (SNVs). The SNVs in a haplotype are considered to be in linkage disequilibrium (LD), which is a measure of the non-random association among variations at different genetic loci (Conrad et al., 2006). An SNV that has been validated by orthogonal approaches or variant frequency is assigned a unique/identifying rs number in the Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al., 1999).

In addition to characterising the genetic complexity due to SNPs and indels, this thesis investigates the distribution and implications of alleles defined by structural variations in *CYP2D6*, *CYP2B6* and *CYP2A6* in African populations. Structural variations are genomic alterations encompassing at least 50 DNA base pairs (bp) (Alkan et al., 2011). CNVs are a subtype of structural variations mainly representing deletions and duplications (Redon et al., 2006). The known structural variations in *CYP2D6*, *CYP2B6* and *CYP2A6* are reviewed in more detail in sections 2.2.3, 2.2.4 and 2.2.5, respectively.

To address the questions above, I have chosen to apply a consensus star allele calling approach involving multiple bioinformatics algorithms for short-read WGS-based investigation of the *CYP2D6*, *CYP2B6* and *CYP2A6* allelic and phenotypic distributions across African populations. This star allele calling approach is supported by an initial benchmarking study (Chapter 3). Short-read data is used because it is more readily available from the H3Africa projects and

collaborations within Africa, and it is the most widely used method of HTS. However, as these genes are challenging to genotype, only samples with high coverage WGS data are included in the analysis. Importantly, this thesis also describes the laboratory validation of novel African-specific alleles found in this study via targeted long-read resequencing.

1.2 Research aims and objectives

The overarching aim of this study is to characterise the allelic variation in the *CYP2D6*, *CYP2B6* and *CYP2A6* pharmacogenes in African populations. A major part of this work involves the benchmarking of existing star allele callers and development of a novel diplotype calling algorithm.

The objectives of this project are;

1. To compare the performance of existing pharmacogene star allele calling algorithms using HTS data.
2. To develop a novel graph- and Nextflow-based algorithm for calling star alleles in highly polymorphic and structurally variant pharmacogenes.
3. To identify common, rare, and potential novel alleles in *CYP2D6*, *CYP2B6*, and *CYP2A6* in African populations using WGS data, and predict the corresponding phenotype distribution.
4. To validate novel African-specific alleles and resolve ambiguous genotypes in *CYP2D6*, *CYP2B6*, and *CYP2A6* by targeted long-read resequencing.

1.3 Thesis outline

This thesis is written in the 'integrated' format recommended by the Wits Faculty of Health Sciences — it contains a literature review, two published articles, two manuscripts intended for future publication, and a discussion of all the papers included in the thesis in the context of the aforementioned aims and objectives. Some repetition is therefore intended in the integrated narrative as the chapters are meant to be self-standing in this format.

Chapter 2 provides a literature review of the pharmacogenomic variation in *CYP2D6*, *CYP2B6* and *CYP2A6* in African versus global populations, and its implications for personalised medicine. Furthermore, this chapter explores the opportunities availed by next-generation technologies, increasing availability of clinical-grade WGS data and the emergence of bioinformatics tools that can phase the plethora of variations in *CYP2D6*, *CYP2B6* and *CYP2A6* using HTS data. This chapter therefore aims to make the case that these emerging resources and opportunities can be used to address the paucity of information about *CYP2D6*, *CYP2B6* and *CYP2A6* pharmacogenomic variation especially in previously unsampled African populations despite the genotyping challenges these genes present.

Chapter 3 presents a published benchmarking study of the existing bioinformatics tools for calling pharmacogene star alleles. The *CYP2D6* gene is used as a model gene for this evaluation given its high importance in drug metabolism and extreme genotyping complexity. This work provides important insights on how the different tools perform on various diplotype combinations and how they genotype individuals with novel alleles which is critical to understand before applying them to novel African genomic data.

Chapter 4 is a published article that describes a novel bioinformatics pipeline (StellarPGx) for calling star alleles in highly polymorphic pharmacogenes. The pipeline was developed to interrogate cytochrome P450 genes initially as these are responsible for ~80% of the metabolism of commonly prescribed medications and they present remarkable genotyping complexity exemplified by *CYP2D6*, *CYP2B6* and *CYP2A6* allelic variation.

Chapters 5 and 6:

The genetic architecture in the *CYP2D6*, *CYP2B6* and *CYP2A6* gene loci is characterised by neighbouring pseudogenes, numerous haplotypes and structural variants as described in chapter 2. Therefore, presenting the results on the allele distributions in these three genes across the African populations represented in this PhD study — not to mention the novel allele descriptions — in the same chapter would be intractable. Given that *CYP2D6* is by far the most polymorphic and widely studied of the three genes, the results from the *CYP2D6* star allele analysis are presented separately (chapter 5), while the *CYP2B6* and *CYP2A6* star allele analyses are presented in chapter 6, albeit for the same study population. Although there is some degree of repetition especially in the methodologies of these two chapters, the clarity gained from presenting *CYP2D6* results in a separate chapter is a key outweighing factor. In addition, CYP2B6 and CYP2A6 are involved in more similar major metabolic pathways (i.e. both are involved in efavirenz and nicotine metabolism), which facilitates discussing them in the same chapter. The empirical studies presented in these two chapters are as follows:

- **Chapter 5** investigates the allelic variation in the *CYP2D6* across African populations using state-of-the-art bioinformatics tools, and provides predicted CYP2D6 metaboliser phenotype distributions which are important given the numerous medications bioactivated by CYP2D6. A subset of the novel African-specific *CYP2D6* alleles and suballeles are further characterised using established long-range PCR strategies and long-read sequencing.
- **Chapter 6** examines the *CYP2B6* and *CYP2A6* pharmacogenomic variation in African populations. CYP2B6 is the key enzyme in the metabolism of the antiretroviral drug efavirenz. *CYP2B6* genetic variation is a major component in the interindividual variability regarding efavirenz response. The variation in *CYP2A6* is of particular importance given its unique impact on nicotine metabolism and smoking behaviours. The results from this study can contribute to a more comprehensive representation of African-specific *CYP2A6*

alleles in the Pharmacogene Variation (PharmVar) Consortium database and ultimately in various genotyping platforms used in clinical and research settings. To the best of my knowledge, this is the first study to apply targeted long-read resequencing in the validation of novel *CYP2A6* alleles.

Chapter 7 presents general conclusions from the thesis, discusses major findings and study limitations, and proposes ideas for future work.

Chapter 2

Literature Review

This chapter provides a brief overview of the burden of disease and adverse drug reactions in Africa, and a more detailed review of the literature on (1) pharmacogenomics and relevant clinical translation approaches (2) *CYP2D6*, *CYP2B6* and *CYP2A6* pharmacogenomic variation in African populations and other biogeographical groups (3) opportunities for catalysing pharmacogenomics and precision medicine in Africa — such as next-generation sequencing technologies, the emergence of star allele calling bioinformatics tools, and the increasing availability of African genomic datasets.

2.1 A snapshot of the burden of disease and adverse drug reactions in Africa

Africa faces a high burden of communicable diseases, not to mention the increasing prevalence of non-communicable diseases (NCDs), with the continent contributing up to around 30% of the global disease burden (Institute for Health Metrics and Evaluation, 2019). NCDs are on the rise in Africa in part due to increasing urbanisation and adopting of Western lifestyles (World Health Organisation, 2020).

As shown by the data generated during the Global Burden of Disease (GBD) study (Vos et al., 2020), the burden of disease — measured as disability-adjusted life years (DALYs) — is not uniform across African geographical regions, and it is certainly different from that in Europe (Figure 2.1). Moreover, the data from the GBD study provides evidence of a worldwide decrease in years of life lost (YLLs) and an increase in years lived with disability (YLDs) between 1990 and 2019. This trend is projected to continue into the future, thus correlating with a shift in health care towards management of chronic disease or disability (Vos et al., 2020). In Africa, health priorities tend towards mitigating the leading causes of DALYs — in particular, the high prevalence of malaria in the tropics, HIV/AIDs in Southern Africa, and cardiovascular diseases in Northern and Southern Africa (Figure 2.1).

A compounding element to the disease burden in Africa is the rapid rise in the burden of adverse drug reactions (ADRs) (Ampadu et al., 2016; Berhe et al., 2015). In particular, the incidence of anti-tuberculosis drug-induced liver injury (DILI), which is often confounded with anti-retroviral DILI, is significant. For example, Petros et al. (2017) and Yimer et al. (2014) reported an incidence of DILI in 11.6% and 17.3%, respectively, in Ethiopian patient cohorts, while Mugusi et al. (2012) reported an 7.8% overall incidence of DILI in a Tanzanian patient

DALY Causes	All Africa	Southern Africa	Central Africa	Eastern Africa	Western Africa	Northern Africa	Europe
Neonatal disorders	1	2	1	1	1	4	18
Enteric infections	2	10	3	3	2	16	23
Malaria	3	27	2	6	3	30	None
Lower respiratory infections	4	5	5	4	4	15	16
Cardiovascular diseases	5	3	4	5	6	1	1
HIV/AIDS	6	1	9	2	8	28	21
Other infectious diseases	7	18	8	7	5	20	25
Neoplasms	8	4	11	9	10	2	2
Congenital birth defects	9	22	13	11	7	10	20
Mental disorders	10	9	10	12	12	3	4
Tuberculosis	11	7	6	8	13	27	24
Nutritional deficiencies	12	19	12	10	9	21	22
Digestive diseases	13	17	14	13	11	13	7
Transport injuries	14	8	7	16	16	6	14
Unintentional injuries	15	13	16	14	14	9	6
Musculoskeletal disorders	16	12	18	17	17	5	3
Neurological disorders	17	14	17	19	15	7	5
Other non-communicable diseases	18	16	21	18	18	11	9
Self-harm and interpersonal violence	19	6	19	15	19	8	12
Chronic respiratory diseases	20	15	20	20	22	14	8
Diabetes mellitus	21	11	22	24	26	12	10
Sense organ diseases	22	20	26	25	24	18	11
Neglected tropical diseases	23	29	15	26	21	26	29
Skin and subcutaneous diseases	24	25	25	22	25	19	15
Maternal disorders	25	28	23	23	23	24	30
Chronic kidney disease	26	21	28	27	27	17	19
STIs (excluding HIV)	27	26	24	21	28	31	31
Hemoglobinopathies and anemias	28	31	27	30	20	25	28
Endocrine, metabolic & immune disorders	29	23	29	29	29	22	17
Substance use disorders	30	24	30	28	30	23	13
Upper respiratory infections	31	30	31	31	32	29	26
Adverse effects of medical treatment	32	32	32	32	31	32	27

Figure 2.1: Burden of disease in Africa in 2019 relative to Europe. the heatmap shows the top 32 causes of disability-adjusted life-years (DALYs) in Africa, relative to Europe. The darker the shading, the more DALYs linked to the disease in a given region. The data about the top 32 DALY causes for countries in each region is available from the Institute for Health Metrics and Evaluation and was primarily collected during the Global Burden of Disease study (Vos et al., 2020).

cohort. In the same vein, other ADRs such as skin and subcutaneous tissue disorders (31.14%), nervous system disorders (17.5%), gastrointestinal disorders (16.1%), respiratory, thoracic and mediastinal disorders (5.7%), blood and lymphatic system disorders (5%) and psychiatric disorders (4.7%) are common ADRs in individual case safety reports across Africa (Ampadu et al., 2016). The factors contributing to the burden of ADRs in various African communities include medication errors, use of substandard medicines, mass treatment programmes, and resource deficiencies in pharmacovigilence systems i.e. ADR detection, assessment, monitoring, and prevention (Ampadu et al., 2016; Ampadu et al., 2018; Isah et al., 2012). However, there are

knowledge gaps as regards to the role of genetic diversity in the inter-individual variability in the efficacy of herbal and western medicines and in the onset of ADRs across African populations (Radouani et al., 2020).

2.2 Pharmacogene variation and clinical relevance across African and global populations

2.2.1 What is pharmacogenomics?

Pharmacogenomics (PGx) is the study of the relationship between genetic variations and individual response to medications. The term 'pharmacogenomics' is often used interchangeably with 'pharmacogenetics'. However, most pharmacogenetics studies are focused on the assessment of the impact of variation in specific genes often involved in drug metabolism (e.g. the cytochrome P450 gene family), while pharmacogenomics involves the study of variation in a wider spectrum of genes in the genome (i.e. including drug target genes, and drug transporters) that contribute to drug response (Nebert, 1999; Scott, 2011; Whirl-Carrillo et al., 2012; Whirl-Carrillo et al., 2021). The genes involved either directly or indirectly in drug response are commonly referred to as 'pharmacogenes' (Gaedigk et al., 2021). Pharmacogenomics research aims to guide physicians to optimise drug therapies by determining the most efficient and safe treatment options for each patient based on their genomic profile (M. Relling and Klein, 2011). Notably, the inter-individual variability in response to standard doses of drugs could also be influenced by age, lifestyle, and environmental factors, but genetic variation is one of the key determinants (Scott, 2011; Meyer et al., 2013).

Pharmacogenomics is therefore an integral part of precision medicine, an emerging approach for disease diagnosis, treatment and prevention that takes into account individual genetic variability alongside environmental and lifestyle factors (Caudle et al., 2016). The precision medicine approach is a contrast to the 'one-size-fits-all' model, where physicians prescribe therapeutics whose safety and efficacy is based on trials in populations that often do not represent the diversity across the globe (Caudle et al., 2016; see also `https://www.pharmgkb.org`). Personalising therapeutic interventions however does not mean making drugs that are unique to a patient, but is rather focused on making evidence-based choices of the best medication and dose for the individual (Whirl-Carrillo et al., 2012).

The concept of precision medicine has been part of health care for a long time, aptly exemplified by blood type matching before blood transfusions (H. G. Klein et al., 2015). Remarkably, even though there has been an increasing hype in recent years, the role of precision medicine in healthcare to date is relatively limited. From the pharmacogenomics perspective, this is in part because a lot of input is required to identify which genetic variants affect drug responses in

individuals from various unsampled populations (Pereira et al., 2021; Radouani et al., 2020; H. Zhang et al., 2019).

As highlighted in the Pharmacogenomics Knowledgebase (PharmGKB, `https://www.pharmgkb.org`), genetic variation affects whether an individual develops an adverse drug reaction (ADR) to a medicine or how well they respond to a medicine by altering a drug's pharmacokinetics (PK) and pharmacodynamics (PD). Pharmacokinetics (PK) entails what the body does to a drug during the four stages of disposition i.e. absorption, distribution, metabolism, and excretion (ADME), while pharmacodynamics refers to the physiological, biochemical and molecular effects of a drug on the body. If a drug is broken down to its active components (bioactivated/ metabolised) too quickly by the body, it might not be as effective, whereas if the drug is metabolised too slowly, the reactive components (metabolites) may build up in the body and cause dangerous ADRs (PharmGKB, 2021). From the metabolism perspective, drug response phenotypes could range from poor to ultrarapid metabolisers depending on the drug administered and germline (heritable) mutations in genes coding for the enzymes involved in the drug's pathway. Similarly, variation in genes encoding drug transporters (e.g. *SLCO1B1* and *ABCB1*) can alter how well drugs enter or exit human body cells, causing the drugs to either not work well or become toxic depending on if their concentration in those cells is decreased or increased. For example, *SLCO1B1* phenotypes range from low to increased function with respect to simvastatin (Ramsey et al., 2014).

Although most of the recent pharmacogenomics work has been focused on ADME genes (da Rocha, Othman, Botha, et al., 2021; Hovelson et al., 2017), the importance of variation in drug targets cannot be overstated. Variation in genes coding for drug target proteins (e.g. VKORC1 and CFTR) could potentially affect the dose of the drug needed to achieve a therapeutic effect by altering the amount of the target protein in the body or by preventing the drug from being able to bind to the structurally altered protein (PharmGKB, 2021). For example, genetic variation that increases or decreases the amount of VKORC1 in the body has been shown to affect the warfarin (anticoagulant) dose required to prevent blood clotting (Johnson et al., 2017).

In recent years, the pharmacogenomics community has categorised pharmacogenes based on importance to prioritise efforts on key biomarkers. The PharmaADME group (`http://www.pharmaadme.org`) classifies ADME genes as "Core" and "Extended", depending on their direct importance or association to drug metabolism, respectively. There are 32 core ADME genes according to this classification and about 260 extended ADME genes. In addition, there are over 80 pharmacogenes involved in the pharmacodynamic response to drugs. Perhaps a more popular initiative is the provision of rigorously curated "Very Important Pharmacogene" (VIP) summaries (`https://www.pharmgkb.org/vips`) by the PharmGKB research community. There are over 60 VIP annotations to date, with varying levels of evidence as regards to their

importance in pharmacogenomics. Alleles in some of the VIP genes are used in drug labelling to aid in treatment decisions based on the genomic profile of a patient. In particular, cytochrome P450 (CYP) genes such as *CYP2D6* (~163 annotations), *CYP2C19* (59 annotations), *CYP2C9* (30 anotations), *CYP3A4* (20 annotations) and *CYP2B6* (10 annotations) have numerous drug label annotations owing to their high polymorphic nature and importance in drug metabolism (`https://api.pharmgkb.org/v1/download/file/data/drugLabels.zip`).

Development and updates to VIP summaries as well as clinical pharmacogenetics implementation guidelines are dependent on pharmacogenomics studies that perform discovery of pharmacogene alleles (alternative gene forms or haplotypes) in various cohorts and/or infer association(s) to drug response phenotypes (PharmGKB gene-specific information tables, 2021). Pharmacogenomics findings are therefore key to catalysing the much-needed widespread implementation of the precision medicine approach. Unfortunately, there is a big disparity in the pharmacogenomics research supporting clinical annotations curated by resources such as PharmGKB in that information based on African populations (and minority groups in the Americas) is rather limited.

2.2.2 Star alleles and phenotype prediction

Given the plethora of SNVs and other forms of genetic variation observed in pharmacogenes in various populations, the pharmacogenomics clinical and research communities rely on an integrated and standardised star ('*') allele nomenclature system as a 'common language' to refer to several variant combinations. The Pharmacogene Variation Consortium (PharmVar, `https://www.pharmvar.org`) and the aforementioned PharmGKB database maintain rigorously curated catalogues of star alleles and relevant population allele frequencies for key ADME genes such as *CYP2D6*, *CYP2C19* and *CYP2B6* among others. Major haplotypes are typically defined by presence of function-altering (core) SNVs and/or structural variants, with a few exceptions such as reference (**1*) alleles (Gaedigk et al., 2021). PharmVar also catalogues a number of suballeles under each major star allele. Suballele status is typically assigned when a newly submitted haplotype has novel variant(s) that do not alter the function of the protein while existing together with known sequence variations that do (Gaedigk, Ingelman-Sundberg, et al., 2018). The star alleles are assigned either a definitive, moderate or limited levels of evidence (`https://www.pharmvar.org/criteria`) to denote whether a haplotype was validated partially or fully i.e. whether all LD SNVs and/or SVs present were ascertained via the sequencing and phasing (inferring haplotypes from variant/sequence data) approaches used. As described in the PharmVar submission criteria, alleles with a 'limited level of evidence' are usually those where exon-only sequencing was done but they could also include WGS-derived alleles e.g. those inferred computationally from challenging loci such as *CYP2B6*.

The activity of these star alleles is curated by the Clinical Pharmacogenetics Implementation

Consortium (CPIC, https://cpicpgx.org; M. V. Relling et al., 2020), with evidence typically from *in vitro* and *in vivo* studies (M. V. Relling et al., 2020). However, computational functionality predictions are also relatively robust for alleles defined by variants with consequences such as stop-gain, stop-loss, and splicing defects. The spectrum of star alleles typically ranges from nonfunctional alleles to increased function alleles — based on CPIC function. However, there is a significant number of star alleles whose activity is unknown/uncertain and these tend to be rare or newly discovered haplotypes (https://www.pharmvar.org). This group also includes alleles where evidence for decrease or increase in the associated protein function is conflicting.

One of the main challenges of implementing pharmacogenomics in the clinic is inferring phenotype from a patient's pharmacogene diplotype (two gene haplotypes, one on each chromosome), as this is key for adjusting therapy (Gaedigk et al., 2017). Given the CPIC functions assigned to star alleles, patient diplotype information can be translated to phenotypes via the activity score (AS) system(s) that have been developed and standardised recently (Gaedigk et al., 2008; Gaedigk, Dinh, et al., 2018; Caudle et al., 2020), albeit for a few genes. For example, CYP2D6 activity scores are currently assigned as shown in Table 2.1 on page 20. Importantly, the AS system also facilitates determining the pharmacogene phenotypic landscape in population-scale pharmaco-genomics studies (McInnes et al., 2021; see also PharmGKB gene-specific information tables, 2021). However, star alleles with unknown/uncertain CPIC function present a hurdle for automated phenotype prediction (Twist et al., 2016). This has sparked the application of predictive approaches such as structural bioinformatics (Borba et al., 2016; Othman et al., 2021) and machine learning (McInnes et al., 2020) to provide insights on the functional impact of these alleles on protein function.

While harmonising genotype to phenotype translation is a major challenge, more clinical and research efforts aimed at pharmacogene variant and haplotype discovery across global populations are also needed to obtain representative pharmacogenomic variation landscapes that would in turn enhance the implementation of precision medicine globally. Previous studies have shown that ADME genes seem to be less sensitive to negative (purifying) selection compared to non-ADME genes since there are more variants per kilobase in ADME genes compared with non-ADME in the respective studied cohorts (Hovelson et al., 2017; da Rocha, Othman, Botha, et al., 2021; J. Li et al., 2011). In particular, the human cytochrome P450 (*CYP*) gene family harbours numerous polymorphisms. Of the 57 CYP genes, 12 (belonging to the *CYP1*, *2*, and *3* families) encode enzymes responsible for 75-80% of all phase one drug metabolism in the liver (Zanger and Schwab, 2013). Characterisation of the variation in key *CYP* genes in African populations could facilitate the implementation of pharmacogenetics-based therapy for a wide range of medications commonly used in Africa.

The next part of this section therefore focuses on highlighting the known variation in three core CYP genes i.e. *CYP2D6*, *CYP2B6* and *CYP2A6* because of their high pharmacogenetic importance in the African context (Alessandrini et al., 2013; Alessandrini and Pepper, 2014; Rajman et al., 2017), allelic diversity and the genotyping challenges they present. These three highly polymorphic pharmacogenes are the focus of the subsequent chapters in this thesis.

2.2.3 *CYP2D6* pharmacogenomic variation

CYP2D6 is one of the most important drug-metabolising enzymes, due to its role in the metabolism of about 25% of clinically prescribed drugs. These drugs include opioids (e.g. codeine – see Figure 2.2), anticancer agents (e.g. Tamoxifen), antidepressants (e.g. amitriptyline and paroxetine), antipsychotics (e.g. risperidone), beta-adrenergic blocking agents (e.g. metoprolol), antiarrhythmics, antiemetics, antihistamines and antiviral agents among others (`https://www.pharmgkb.org`; see also Flockhart et al., 2021).

2.2.3.1 Genomic architecture of the *CYP2D* region

The *CYP2D6* gene is located on the chromosome 22q13.1 locus in the human genome, neighbouring two highly similar pseudogenes i.e. *CYP2D7* and *CYP2D8P* (Kimura et al., 1989; Heim and Meyer, 1992) (see Figure 2.3 on page 18). *CYP2D7* has a T-insertion in exon one, thus rendering the gene nonfunctional, while *CYP2D8* is known to have accumulated several gene-disrupting mutations (Nofziger et al., 2020). The high homology (~94%) between *CYP2D6* and *CYP2D7*, and the presence of highly similar repetitive sequences (i.e. REP6 and REP7 respectively) make the CYP2D region prone to complex structural rearrangements (see Figure 2.4 b-e on page 19), and presents difficulties for PCR and sequencing. *CYP2D6*-specific and *CYP2D7*-specific XL-PCR primers are difficult to design given the aforementioned high homology between the gene units, while the non-specific XL-PCR primers can lead to erroneous allele characterisation, especially for samples with hybrid alleles, due to co-amplification of *CYP2D6* and *CYP2D7* regions (Gaedigk, 2013; Nofziger and Paulmichl, 2018). In particular, *CYP2D6* targeted sequencing approaches are prone to missing structural variant alleles or leading to inaccurate allele assignments (Turner et al., 2021). Furthermore, read misalignments may occur between *CYP2D6* and *CYP2D7* (or *CYP2D8*) especially when using short-read sequence data, thus leading to variant calling errors and inconsistent CNV detection (Yang et al., 2017).

PharmVar has catalogued over 130 different *CYP2D6* star alleles, evidencing how variable the gene is (`https://www.pharmvar.org/gene/CYP2D6`; Accessed: Aug 5, 2021). A number of these star alleles are known to result in the production of a nonfunctional protein, or to decreased or increased *CYP2D6* metabolising activity (Figure 2.3 on page 18). The CPIC function of star alleles is usually ascertained using selective probe drugs including sparteine,

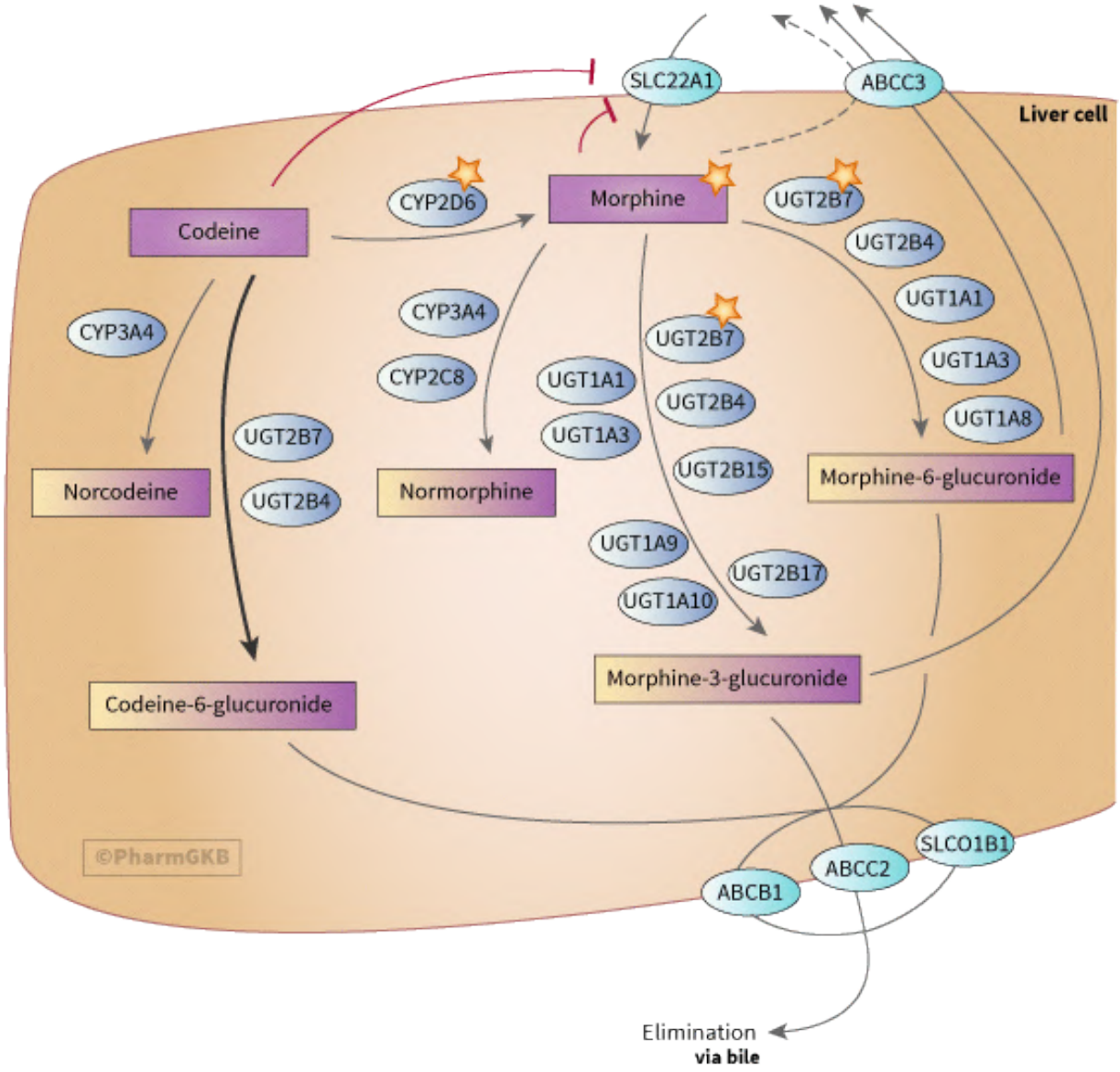


Figure 2.2: Schematic representation of the candidate genes involved in metabolism of codeine and morphine, which are commonly prescribed for pain relief. This image was obtained from PharmGKB (`https://www.pharmgkb.org/pathway/PA146123006`) under the Creative Commons BY-SA 4.0 license. CYP2D6 is responsible for the conversion of up to 15% of the codeine intake to morphine, which is much more active compared to norcodeine (CYP3A4-mediated metabolism) and codeine-6-glucuronide (UGT2B7-mediated metabolism). CYP2D6 poor metabolisers are unable to convert codeine to morphine efficiently and may fail to experience pain relief as a result — when given the standard dose. Ultrarapid metabolisers may metabolise codeine too quickly thus causing morphine intoxication. Notably, 80% of variant annotations on codeine involve *CYP2D6*. ABCC, ATP binding cassette subfamily C; ABCB, ATP binding cassette subfamily B; CYP, cytochrome P450; SLCO, solute carrier organic anion transporter; UGT, UDP-glucuronosyltransferase

---

debrisoquine and metoprolol (Nofziger et al., 2020).

2.2.3.2 *CYP2D6* allele and phenotype diversity

There are varying *CYP2D6* allele frequencies within and between populations (Gaedigk et al., 2017; Rajman et al., 2017; Y. Zhou et al., 2017). This is in part responsible for differences in the distribution of CYP2D6 metaboliser phenotypes (Gaedigk et al., 2017).

The *CYP2D6* poor metaboliser phenotype is only possible when an individual has two nonfunctional *CYP2D6* haplotypes (Table 2.1 on page 20). *CYP2D6*4*, which is defined by the function-obliterating splice acceptor variant (1847G>A, rs3892097), is one of the more common nonfunctional *CYP2D6* star alleles. This allele is most prevalent in Europeans (AF = 18-25%) (Gaedigk et al., 2017), intermediate in SSA (1.9-13%) (Gaedigk et al., 2017; Matimba et al., 2009; Wright et al., 2010), Egyptians (AF = ~9.6%) (Mutawi et al., 2021), African Americans (AF = ~4.8%) and Central/South Asians (AF = ~9%), but rare among East Asians (AF = ~0.5%) (`https://www.pharmgkb.org/page/cyp2d6RefMaterials`; Y. Zhou et al., 2017). These frequency differences imply that whereas *CYP2D6*4* may account for most poor metaboliser phenotypes (dependent on whether the second allele is also null) in Europeans, that is not necessarily the case in African and Asian populations. Other notable nonfunctional alleles include the *CYP2D6* gene deletion (*CYP2D6*5*), which is present at a frequency of 1.5 to 5% in most populations and notably with a higher frequency in SSA (~5%) compared to Europeans (~3%), and *CYP2D6*3*, defined by a frameshift variant (2550delA, rs35742686) which is mainly in Europeans (PharmGKB and CPIC *CYP2D6* reference materials, 2021). These differences in the frequencies of nonfunctional *CYP2D6* alleles underscore the need for clinical pharmacogenetic implementation guidelines based on population-specific allele information.

The *CYP2D6* intermediate metabolisers (Table 2.1 on page 20) are another important group to consider in order to adjust the choice and/or dosage of drugs metabolised by *CYP2D6* accordingly. The *CYP2D6* IM phenotype typically arises when an individual has either two decreased function haplotypes or one decreased function haplotype alongside a nonfunctional haplotype (Caudle et al., 2020). Two of the decreased function star alleles i.e. *CYP2D6*17* (defined by 1022C>T, rs28371706) and *CYP2D6*29* (defined by 1660G>A, rs61736512 and 3184G>A, rs59421388) are remarkably African-specific and have been shown to be largely responsible for the IM phenotype in African populations (Gaedigk et al., 2017; Masimirembwa et al., 1996; Wennerholm et al., 2001; Wennerholm et al., 2002). From the studies to date, *CYP2D6*17* is present at frequencies of ~20% in SSA, ~3% in northern Africans and ~17% in African Americans, while the frequency in European and Asian populations is less than 0.5%. *CYP2D6*29* is present at frequencies of ~12% in SSA and ~8% in African Americans, and it is virtually non-existent in individuals of European and Asian ancestry (PharmGKB and CPIC *CYP2D6* reference materials, 2021). Conversely, *CYP2D6*10* is not as common in African populations (AF = ~6%), and is even less common in Europeans (AF = ~1.5%) yet it occurs at frequencies

ranging from 40–60% in East Asians (PharmGKB and CPIC *CYP2D6* reference materials, 2021; see also Y. Zhou et al., 2017). Other *CYP2D6* decreased function alleles that have been shown to impact the enzyme's oxidation capacity include *CYP2D6*41* and *CYP2D6*9* to mention but a few (Nofziger et al., 2020).

As mentioned earlier, there are a number of structural variations in the *CYP2D* locus and these include *CYP2D6* CNVs (gene deletion and duplication/multiplications), *CYP2D6-2D7* and *CYP2D7-2D6* hybrid genes, and complex tandem rearrangements (Figure 2.4 on page 19). Various studies have described *CYP2D6* gene duplications occuring for **1, *2, *4, *6, *9, *10, *17, *29, *35, *41, *43, and *45* among other alleles (Gaedigk et al., 1991; Gaedigk et al., 2007; Gaedigk et al., 2012; Steen et al., 1995; see also PharmVar structural variation for *CYP2D6*, 2021). It is important to note that phasing duplications (determining which of the alleles is duplicated/multiplicated) is non-trivial due to the shear amount of variants in the *CYP2D* region, some of which may be shared among haplotypes. Phasing these duplications is important because not all of them involve functional gene copies i.e. a duplicated allele is not always an increased function allele. For example, the *CYP2D6*4xN* alleles are all nonfunctional (AS = 0, since **4* is nonfunctional), while *CYP2D6*17x2* and *CYP2D6*29x2* are both normal function alleles (AS = 2, since **17* and **29* are decreased function alleles with AS = 1). It goes without saying that only duplications involving normal function alleles result in increased function alleles (e.g. *CYP2D6*1xN*, *CYP2D6*2xN* and *CYP2D6*45xN*). The activity value for these alleles depends on how many copies are present (AS = 1xN) (Caudle et al., 2020). The frequency of these CNVs varies considerably. For example, the *CYP2D6*1xN* allele is most prominent in the Oceanian population (AF = ~12%) but it has so far been shown to be less common in SSA (AF= ~1.1%). It is remarkable that SSA populations (AF = ~1.5%) and African American populations (AF = ~2.6%) have a higher frequency of **4xN* compared to Europeans (~0.7%) yet the latter have a far greater frequency of the non-functional **4* allele as aforementioned (`https://www.pharmgkb.org/page/cyp2d6RefMaterials`). However, caution needs to be taken while interpreting these frequencies given the aforementioned genotyping challenges in this region and potential mistyping due to the presence of hybrid gene copies.

Hybrid genes comprising portions of *CYP2D6* and *CYP2D7* are believed to be a result of unequal crossover events. *CYP2D7-2D6* hybrids characteristically comprise at least a *CYP2D7*-derived exon 1 (rendering them nonfunctional due to an inherent T-insertion) and a *CYP2D6*-derived 3' portion (Gaedigk et al., 2010) (Figure 2.4). Multiple *CYP2D7-2D6* hybrids exist due to differences in their switch regions from *CYP2D7* to *CYP2D6*. However, since *CYP2D7-2D6* hybrids are all non-functional, they are consolidated under a single star allele designation i.e. *CYP2D6*13* (`https://www.pharmvar.org/gene/CYP2D6`). Conversely, *CYP2D6-2D6* hybrids comprise a *CYP2D6*-derived 5'-portion which then switches to *CYP2D7* (various switch regions have been discovered). Some *CYP2D6-2D7* hybrids (category A) have a fully

CYP2D7-like downstream region with the 1.7 kb long 'spacer sequence' (see Figure 2.4). However, in some cases (category B), the 3'-portions of the *CYP2D6-2D7* hybrids switch back to *CYP2D6*, thus ending in a *CYP2D6*-like downstream region (REP6) without the spacer sequence (Black et al., 2012; Gaedigk et al., 2012; Kramer et al., 2009; see also PharmVar structural variation for *CYP2D6*, 2021). It is unclear as to whether all *CYP2D6-2D7* hybrids are nonfunctional at this point since the common exon 9 conversion has not yet been fully functionally characterised to determine the extent to which it decreases or obliterates the function of the resulting enzyme (PharmVar structural variation for *CYP2D6*, 2021). These hybrid alleles are important to consider not only due to their functional impact, but also because they could affect the accuracy of genotyping assays thus leading to incorrect phenotype predictions (Turner et al., 2021). For example, category B *CYP2D6-2D7* singletons may support amplification with certain XL-PCR-based *CYP2D6* gene deletion (**5*) assays thus leading to false-positive *CYP2D6*5* calls (PharmVar structural variation for *CYP2D6*, 2021).

In addition to the typical aforementioned hybrid genes, other *CYP2D7* conversions have been shown to occur within *CYP2D6*. The well-characterised conversions to date in this category include: the benign intron 1 conversion found in many alleles with a *CYP2D6*2* background, the 5'-UTR conversion (benign) in *CYP2D6*35.002*, and the exon 2 conversion in *CYP2D6*82* (unknown function) (`https://www.pharmvar.org/gene/CYP2D6`).

The presence of tandem arrangements perhaps epitomises the complexity associated with investigating genetic variations in the *CYP2D* region. Whereas gene units in duplications and multiplications are identical, tandem arrangements comprise gene units that are non-identical, and they may harbor two or more gene copies. More often than not, at least one gene copy in these tandem arrangements is a *CYP2D7-2D6* or *CYP2D6-2D7* hybrid (Figure 2.4e) (Gaedigk et al., 2010; Gaedigk, Turner, et al., 2019; Hosono et al., 2009; see also PharmVar structural variation for *CYP2D6*, 2021). The distribution of these tandems varies and is challenging to infer accurately from non-orthogonal genotyping approaches. *CYP2D6*36+*10* is one of the well-characterised tandems and it found predominantly in individuals of East Asian ancestry while the *CYP2D6*68+*4* tandem is frequently found in Europeans (`https://www.pharmgkb.org`). In rare cases, *CYP2D6*17x2* presents as a tandem i.e. when the duplicated copy contains a CYP2D7-like downstream region (`https://www.pharmvar.org/gene/CYP2D6`).

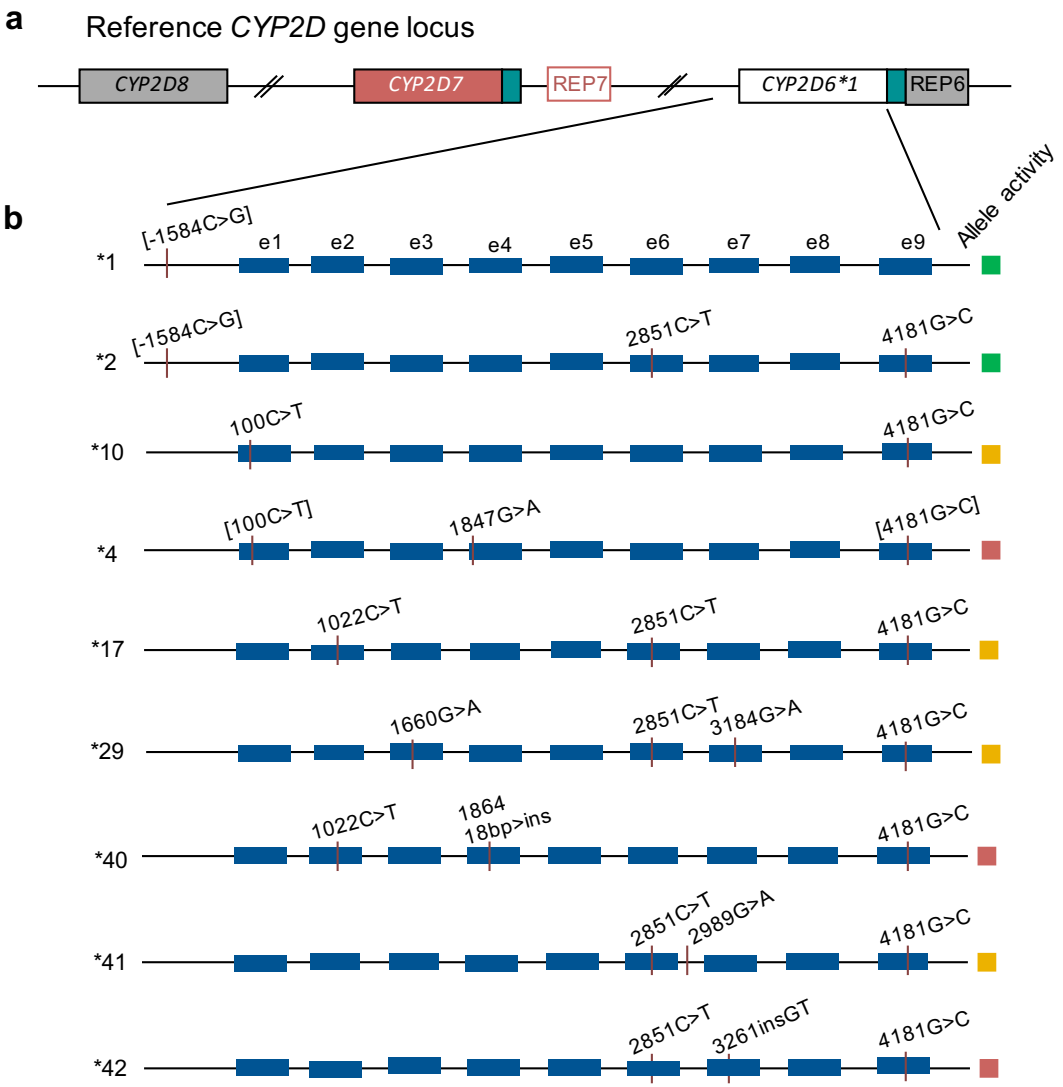


Figure 2.3: Overview of the *CYP2D6* SNV-defined star alleles (haplotypes). (**a**) Schematic representation of the reference *CYP2D* locus harbouring *CYP2D6* and the neighbouring nonfunctional pseudogenes, *CYP2D7* and *CYP2D8*. The locus is located on chromosome 22 and is shown here in reverse since it occurs on the minus strand. In particular, *CYP2D7* is highly similar to *CYP2D6* (~94% homology). REP6 and REP7 comprise repetitive segments that are common to both *CYP2D6* and *CYP2D7*. Notably, *CYP2D7* features a 'spacer sequence' in its downstream region i.e. between REP7 and the adjacent teal-coloured box. The teal-coloured boxes indicate identical unique sequences downstream of *CYP2D6* and *CYP2D7*. (**b**) Examples of major star alleles and their core SNVs. Note that these alleles might habour additional (benign) SNVs that are not shown. The square '[]' brackets on some variants indicate that those variants needn't be present for the assignment of the allele where they occur (i.e. one or more of the other core variants in LD is responsible for the observed allele activity). *CYP2D6*1* and *CYP2D6*2* represent the main backbones during allele assignment i.e. they are considered as default assignments when other core SNVs are absent — presence of 2851C>T distinguishes *CYP2D6*2*-like alleles from *CYP2D6*1*-like alleles. *CYP2D6*17*, **29*, **40* and **42* are African-specific as highlighted in section 2.2.3. The evidence-based CPIC function of these alleles is colour-coded on the right as follows; green, normal function; yellow, decreased function; red, no function. Activity values of 1, 0.5 and 0 are assigned, respectively, based on these activity categories (except **10* which is assigned a value of 0.25). The information for this figure is originally reviewed by Gaedigk (2013), Twist et al. (2016) and Nofziger et al. (2020).

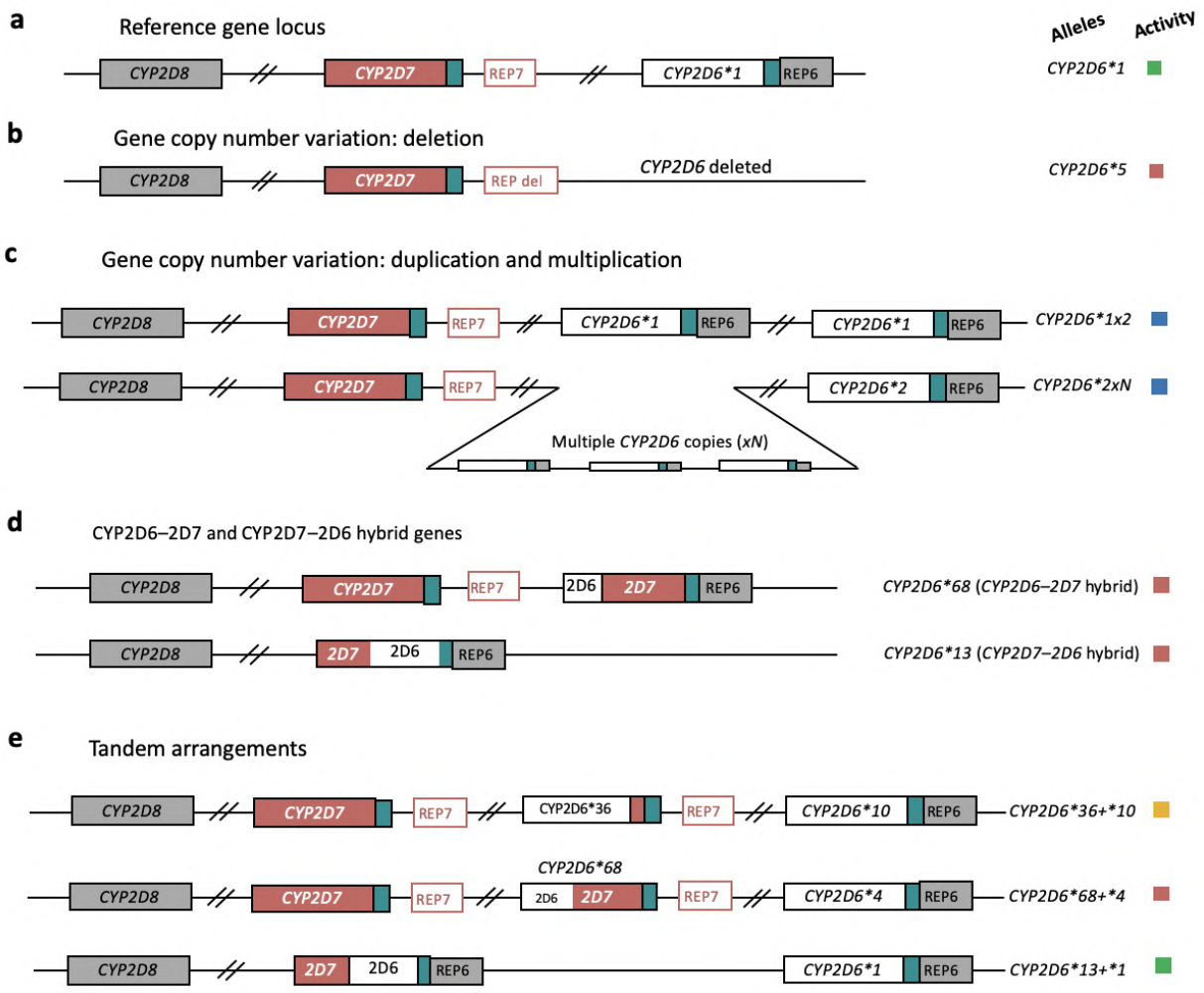


Figure 2.4: Overview of the structural variations in the *CYP2D* locus. (**a**) Representation of the *CYP2D* reference locus comprising the *CYP2D6*1* haplotype and the two paralogous nonfunctional pseudogenes, *CYP2D7* and *CYP2D8*. (**b**) Shows the *CYP2D* locus after a large deletion event. The entire *CYP2D6* gene is deleted in the resulting allele (*CYP2D6*5*), which is also characterised by the fusion of REP6 and REP7 segments to form the REPdel element. (**c**) Illustration of gene duplication and multiplication events. *CYP2D6*1xN* has increased function – the activity value is 1 x number of copies (N). For *CYP2D6*2xN* (increased function), the number of duplicate copies ranges from one to 12. Not all duplications or multiplications have increased function e.g. *CYP2D6*29x2* has normal function while *CYP2D6*4xN* is nonfunctional. (**d**) Depiction of hybrid genes composed of *CYP2D7* and *CYP2D6* portions. *CYP2D6-2D7* hybrids have a *CYP2D6*-derived 5'-end and a *CYP2D7*-derived 3'-end. A few *CYP2D6-2D7* hybrids can occur singly e.g. **68* (shown here) but they most commonly occur in tandem arrangements. Conversely, *CYP2D7-2D6* hybrids have a *CYP2D7*-derived 5'-end and a *CYP2D6*-derived 3'-end. Even though multiple switch regions have been characterised for *CYP2D7-2D6* hybrids, they are all consolidated as *CYP2D6*13* given that they share a function-obliterating T-insertion in exon 1. (**e**) Visualisation of example tandem arrangements which are characterised by two or more copies of CYP2D6, some of which are hybrid alleles. The legend for colour-coded allele activity is as follows; green, normal function; yellow, decreased function; red, no function; blue, increased function. The information for this figure is originally reviewed by Gaedigk (2013), Twist et al. (2016), Nofziger et al. (2020) and is also summarised in the PharmVar structural variation for *CYP2D6* (2021) document.

Generally, it is clear that accurately genotyping *CYP2D6* is challenging as aptly described by Gaedigk (2013) and Nofziger and Paulmichl (2018). Aside from the probable coamplification of pseudogenes, erroneous variant calling and difficult-to-resolve recombination events (especially with short-read data), the potential existence of uncharacterised population-specific variations that could affect the accuracy of commonly used genotyping assays is another big challenge (Turner et al., 2021). Therefore, using high-coverage WGS data would be preferable for large-scale research involving individuals from understudied populations so as to confidently investigate rare and novel high-impact variations and also to prioritise haplotypes that need follow-up with more comprehensive genotyping approaches.

2.2.3.3 *CYP2D6* genotype to phenotype translation

Table 2.1: CPIC and DPWG consensus recommendations for *CYP2D6* genotype to phenotype translation

Haplotypes					
CPIC clinical function	No function	Decreased	Normal	Increased	Uncertain/unknown
Activity value	0	0.5 (0.25 for *10)	1	>1	Indeterminate
Diplotypes					
Phenotype	PM	IM	NM	UM	Indeterminate
Activity score	0	0.25 - 1	1.25 - 2.25	>2.25	Indeterminate

PM, poor metaboliser; IM, intermediate metaboliser; NM, normal metaboliser; UM, ultrarapid metaboliser; CPIC, Clinical Pharmacogenetics Implementation Consortium; DPWG, Dutch Pharmacogenetics Working Group
This summary is adapted from (Caudle et al., 2020)

Given the variability in *CYP2D6*, CPIC guidelines have been published for multiple *CYP2D6* gene-drug pairs to provide evidence-based therapeutic recommendations (e.g. for tamoxifen, opioids, and antidepressants) based on genotype (`https://cpicpgx.org/guidelines`). CPIC and other clinical pharmacogenomics consortia also realised the need to standardise the translation of *CYP2D6* genotype to phenotype so as to aid in decision making involving treatment regimens (Caudle et al., 2020). The consensus recommendations on how to apply the *CYP2D6* activity score used are summarised in Table 2.1. However, there are important caveats to consider regarding *CYP2D6* genotype-phenotype translation via the AS system.

Firstly, *CYP2D6* metaboliser phenotypes could manifest in a substrate-specific manner (Gaedigk, Dinh, et al., 2018). For example, *CYP2D6*10* is known to exhibit varying levels of decreased function toward a range of substrates (Caudle et al., 2020). Perhaps more relevant in the African context, *CYP2D6*17*, which exhibits reduced activity towards tamoxifen, has been shown to metabolise risperidone at a normal and sometimes increased rate (S.-F. Zhou, 2009; Cai et al., 2006). A somewhat similar challenge is the difficulty in translating the genotype's impact

on the metabolism of drugs which could be bioactivated via alternative pathways. This is accentuated for CYP substrates as most of these are metabolised and/or cleared via by multiple competing pathways (von Moltke et al., 1998). The percentage contribution of each enzyme to the metabolism process is obviously a factor. These confounding variables underscore the need for caution when extrapolating allele-specific functional data from one drug or substrate to another (Nofziger et al., 2020).

Secondly, there are genetic factors (other than the *CYP2D6* diplotype) that could impact mRNA and the level of *CYP2D6* expression. For example, a long-range enhancer SNV (rs5758550) in partial LD with the **2*-defining variant (rs16947, R296C) has been shown to confer up to a 2-fold increase in *CYP2D6* transcription levels (D. Wang et al., 2014). However more studies are needed to fully ascertain the functional impact of this long-range enhancer SNV (found over 100kb downstream of *CYP2D6*) especially given that it sometimes occurs in LD with other haplotypes with varying functions e.g. **1* (normal function), **5* (no function), and **41* (decreased function) (Gaedigk, Dinh, et al., 2018). Furthermore, *CYP2D6* expression levels could be modulated via transcription factors (S. S. Lee et al., 2008) and miRNA (Zeng et al., 2017).

Thirdly, it is important to note that *CYP2D6* is prone to inhibition by a number of agents e.g. methadone (Gelston et al., 2012), paroxetine and fluoxetine (Bahar et al., 2018). Some of these inhibitors alter the pharmacokinetic profile of drugs metabolised by *CYP2D6* thus resulting in phenoconversion i.e. causing an individual to present with a metaboliser phenotype contrary to that predicted from their genotype due to inhibition of the metabolising enzyme (Shah and Smith, 2015). *CYP2D6* inhibitors could affect the accuracy of *in vivo* genotype–phenotype correlations particularly if participants do not fully disclose their medication history (Bahar et al., 2018). Interestingly, some components of herbal remedies which are metabolised by *CYP2D6* could also interfere with the metabolism of the prescribed therapeutics or probe drugs in PK studies. For example, Thomford et al. (2018) showed that extracts from *Hyptis suaveolens*, a bush mint used in many African countries (and beyond) to treat diseases such as malaria and eczema, inhibit *CYP2D6* as well as *CYPs 1A2* and *3A4*. Therefore, drug-drug and herbal-drug interactions have the potential to unfavourably interfere with Western medications.

Perhaps the more concerning caveat with using the AS system is the aforementioned relatively large number of alleles with unknown/uncertain function, particularly for *CYP2D6*. Some of these alleles e.g. *CYP2D6*43*, **73* and **74* have been found in African populations (Wright et al., 2010) and they cannot be assigned AS values at the moment, meaning that diplotypes with these alleles would be assigned an indeterminate metaboliser phenotype. More studies are therefore warranted to provide insight on the functional consequences of such alleles.

2.2.4 *CYP2B6* pharmacogenomic variation

CYP2B6, which is mainly expressed in the liver, is the only gene in the human *CYP2B* subfamily that encodes a functional enzyme (Desta et al., 2021; H. Wang and Tompkins, 2008). *CYP2B6* has been extensively studied in pharmacogenetics due to its central role in the metabolism and clearance of efavirenz, a key drug for the treatment of HIV type-1 infection (Figure 2.5) (Dhoro et al., 2015; Nyakutira et al., 2008; see also `https://www.pharmgkb.org/chemical/PA449441/clinicalAnnotation`). Notably, patients categorised as poor metabolisers are at increased risk of central nervous system adverse events if given the standard dose of efavirenz (600 mg/day). It is estimated that *CYP2B6* is involved in the bioactivation of 8–13% of clinically prescribed drugs (H. Wang and Tompkins, 2008; Zanger et al., 2007). Other than efavirenz, some of the key medications where *CYP2B6* genetic variation and metabolic status is pertinent to the metabolism and/or toxicity include methadone (opioid agonist and analgesic), ketamine (antidepressant), and bupropion (antidepressant) (Zanger et al., 2007).

2.2.4.1 Genomic architecture of the *CYP2B* region

The *CYP2B6* gene is located on chromosome 19 (19q13.2) near a highly homologous nonfunctional pseudogene *CYP2B7P* (see Figure 2.6) within the large *CYP2ABFGST* gene cluster (Hoffman et al., 2001). There is a 96% homology between the exonic sequences of *CYP2B6* and *CYP2B7* (i.e. they differ at only 60 nucleotide positions) (Desta et al., 2021). The presence of the *CYP2B7* pseudogene neighbouring *CYP2B6* therefore presents genotyping challenges such as off-target amplification or sequencing, short-read misalignment and/or error-prone variant detection.

2.2.4.2 *CYP2B6* allele and phenotype diversity

PharmVar has catalogued 38 *CYP2B6* core alleles to date with either definitive, moderate or limited level of evidence (`https://www.pharmvar.org/gene/CYP2B6`). *CYP2B6*1* is a default assignment when core SNVs genotyped by various platforms are missing in a sample. For this reason, the frequency of **1* tends to be inflated. *CYP2B6*6*, a decreased function star allele which is defined by the Q172H and K262R amino acid changes (Figure 2.6), is the most common non-reference star allele in nearly all superpopulations (`https://www.pharmgkb.org/page/cyp2b6RefMaterials`). This allele is prominent in SSA, African American and Oceanian populations. The high *CYP2B6*6* frequency in SSA (up to ~37%) has been shown to be in part responsible for the significant proportion of efavirenz poor metaboliser phenotypes in SSA (see Figure 2.7 on page 26) based on studies on Zimbabwean and other populations (Dhoro et al., 2015; Maimbo et al., 2012; Nyakutira et al., 2008; Peko et al., 2019). The 15631G>T (rs3745274, Q172H) polymorphism is intriguing as its diminishing impact on *CYP2B6* activity results mainly from causing incorrect splicing of the *CYP2B6* pre-mRNA (thus

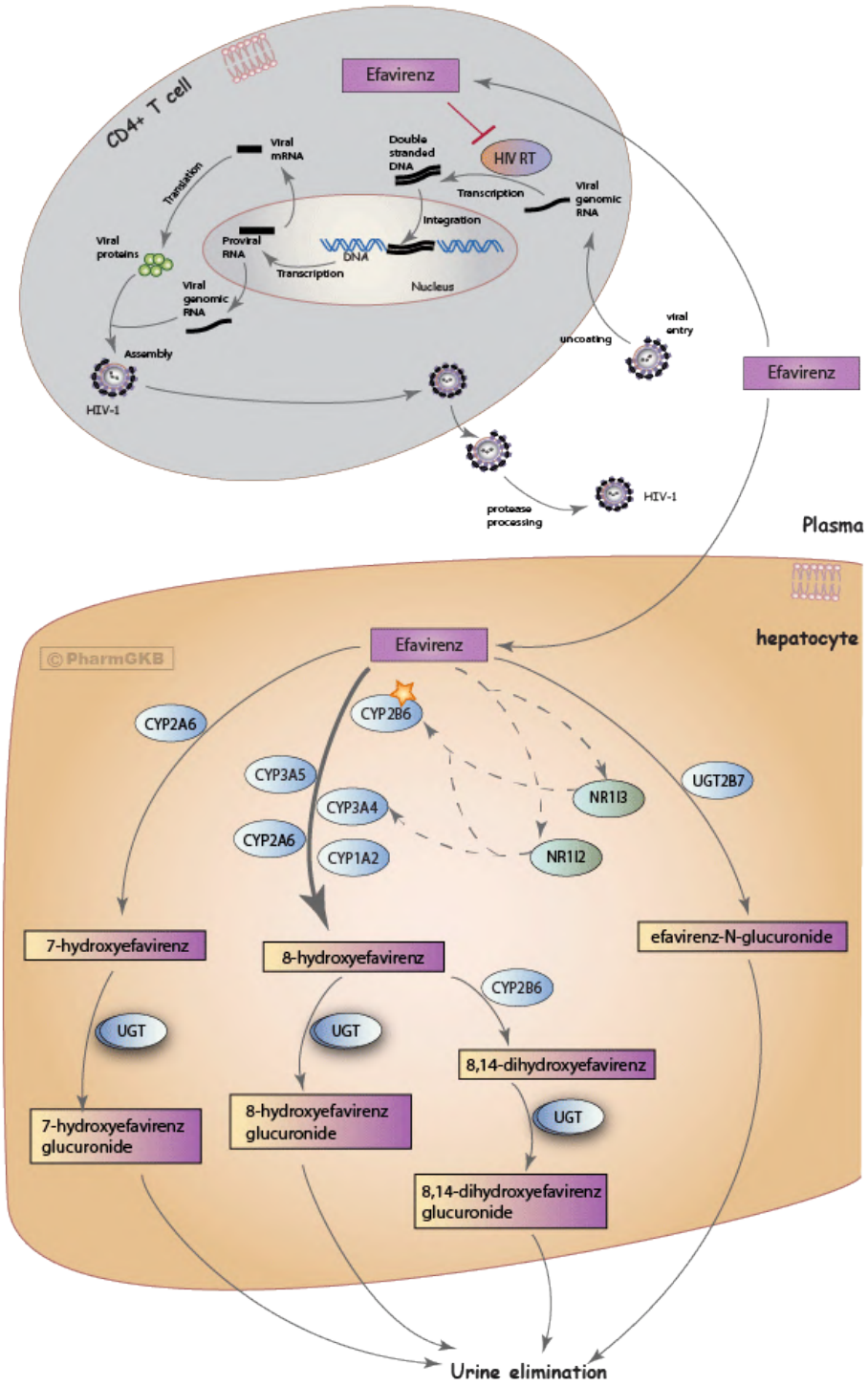


Figure 2.5: Schematic representation of the mechanism of action against HIV, and the candidate genes involved in metabolism of efavirenz. CYP2B6 catalyses the formation of 8-hydroxyefavirenz (8-hydroxy-EFV) i.e. the major route of efavirenz metabolism. The formation rate of 8-hydroxy-EFV varies considerably mainly due to *CYP2B6* genetic variation, and may also be influenced by variations in *CYP2A6*, *CYP1A2*, *CYP3A4* and *CYP3A5* (Ogburn et al., 2010). CYP2B6 poor metabolisers (e.g. patients with the **6/*6* genotype) taking efavirenz are at increased risk of adverse events such as neurologic toxicity. This image is reproduced from PharmGKB (`https://www.pharmgkb.org/pathway/PA166123135`) and is licensed under a Creative Commons BY-SA 4.0 license. CYP, cytochrome P450; NR1, Transcription factors pregnane X receptor (PXR, NR1I2) and constitutive androstane receptor (CAR, NR1I3); RT, reverse transcriptase; UGT, UDP-glucuronosyltransferase.

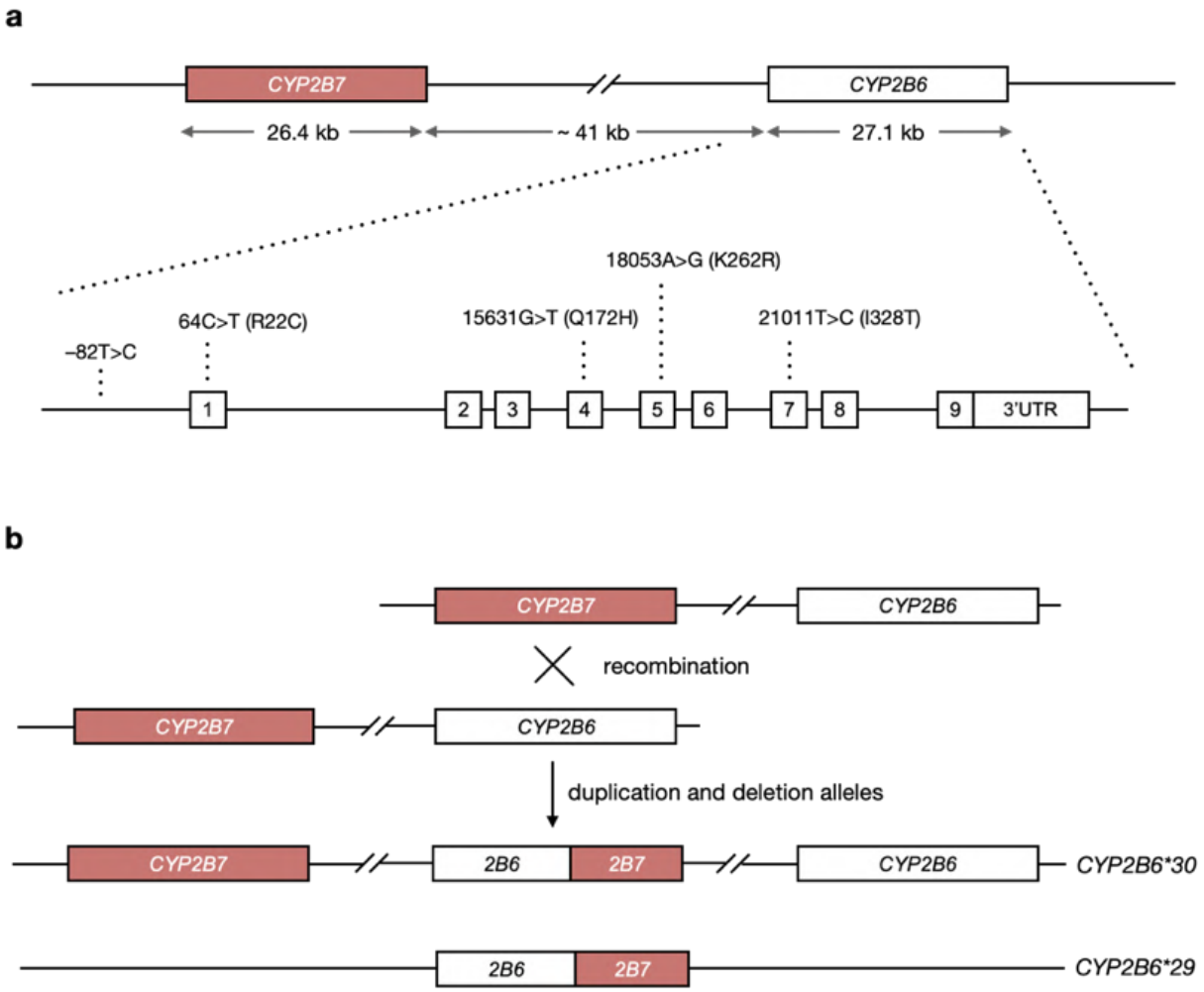


Figure 2.6: Graphical overview of the *CYP2B* locus and allelic variation. (**a**) Schematic representation of the *CYP2B* gene locus comprising *CYP2B6* and the neighbouring highly similar pseudogene, *CYP2B7* on the forward strand of chromosome 19. The enlarged portion shows the location of selected *CYP2B6* SNVs relative to the ATG-start on the RefSeqGene LRG_1267. The -82T>C (**22*) is a promoter variant associated with increased function. It is also found in other alleles. The 64C>T (p.R22C) variant defines **2*. 15631G>T (p.Q172H) and 18053A>G (p.K262R) define **9* (decreased function) and **4* (increased function), respectively. Importantly, these two variants occur together in *CYP2B6*6*, a decreased function allele that is prevalent in African populations. 21011T>C (p.I328T) defines **18* (nonfunctional). Notably, these variants can occur in other haplotypes with varied functions. (**b**) Depicts a reciprocal recombination event resulting in the formation of *CYP2B6*29* (partial deletion) and *CYP2B6*30* (duplication event). *CYP2B6*29* has been characterised as a decreased function allele owing to the 23 amino acid changes it harbours, while *CYP2B6*30* harbours 15 amino acid changes and is nonfunctional. The high intron 4 homology between *CYP2B6* and *CYP2B6* facilitates the recombination thus resulting in these two hybrids with portions from *CYP2B6* and *CYP2B7* and characteristic intron 4 switch regions. The information for this figure is originally reviewed by (Desta et al., 2021). The information on *CYP2B6* structural variants is also summarised in the PharmVar database (`https://www.pharmvar.org/gene/CYP2B6`).

leading to a shorter mRNA lacking exons 4 to 6) rather than the deleteriousness of the amino acid change itself (Hofmann et al., 2008). Importantly, the 15631G>T variant is also found in combination with other variants in multiple alleles including **7, *9, *13, *19, *20, *26, *34, *36,* and **38* (`https://www.pharmvar.org/gene/CYP2B6`). The second **6* core variant 18053A>G (K262R) is the defining variant of **4* and is also found in **13*, **18.002* and **19* to mention but a few. These scenarios where multiple SNVs reside in more that one star allele pose a challenge in genotyping (Desta et al., 2021) and therefore frequency estimations may be inaccurate as a result depending on the genotyping platform used.

In the African context in particular, another key functionally defective *CYP2B6* allele is *CYP2B6*18* – defined by 21011T>C (I328T) (Dhoro et al., 2015; Röhrich et al., 2016). This nonfunctional allele is present at a frequency averaging ~5.8% in SSA and ~3.3% in African Americans/Afro-Caribbeans while it seems to be non-existent in European, East Asian and South Asian populations (`https://www.pharmgkb.org/page/cyp2b6RefMaterials`). Individuals with the *CYP2B6*6/*18*, *CYP2B6*6/*6*, and *CYP2B6*18/*18* diplotypes are all assigned a PM status with respect to efavirenz treatment (Table 2.2) (Desta et al., 2019).

In addition to the SNV-defined alleles, structural variations have been discovered in the *CYP2B6* gene region. These include the *CYP2B6*29* (Rotger et al., 2007; Martis et al., 2013) and *CYP2B6*30* (Martis et al., 2013) hybrid alleles (see Figure 2.6). Both of these hybrids are believed to be a result of unequal crossover recombination events – exchange of genetic material between maternal and paternal chromosomes – facilitated by Alu elements within the intron 4 portion of both *CYP2B6* and *CYP2B7*. The *CYP2B6*29* allele has a *CYP2B7*-derived 5'-portion and a *CYP2B6*-derived 3'-portion (hence its classification as a *CYP2B7-2B6* hybrid), suggesting that it is possibly a product of a large deletion between *CYP2B6* and *CYP2B7* (Desta et al., 2021). This deletion allele is therefore not completely analogous to *CYP2D6*5*, where the entire *CYP2D6* gene is deleted. However, *CYP2B6*29* has 23 amino-acid changes in its open reading frame and is currently annotated as a decreased function allele (Rotger et al., 2007). Conversely, the *CYP2B6*30* allele has a *CYP2B6*-derived 5'-portion and a *CYP2B7*-derived 3'-portion, hence its classification as a *CYP2B6-2B7* hybrid (PharmVar structural variation for *CYP2B6*, 2021). *CYP2B6*30* (nonfunctional) has been shown to occur as an additional gene copy located between *CYP2B7* and a non-hybrid *CYP2B6* allele (Martis et al., 2013) (Figure 2.6). It is believed that *CYP2B6*30* is a product of an unequal cross-over and duplication event (Desta et al., 2021). These structural variations are important to consider because not only do they diminish the function of the CYP2B6 enzyme but they could also affect the accuracy of *CYP2B6* genotyping platforms.

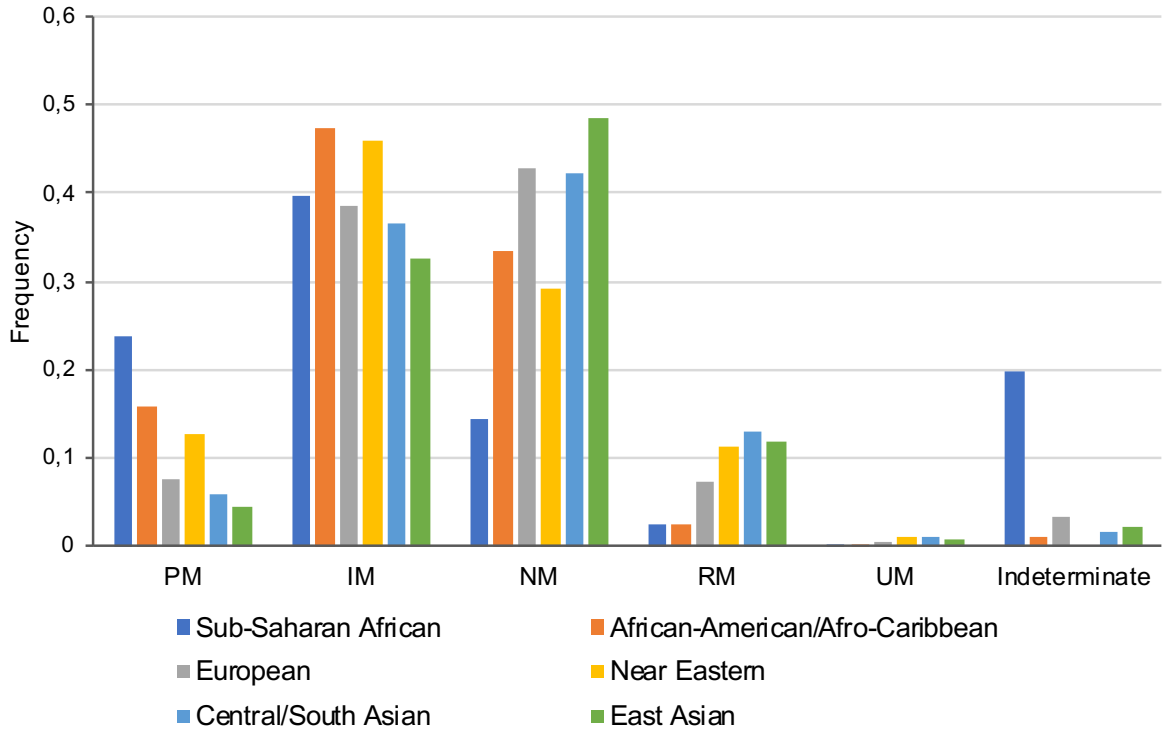


Figure 2.7: CYP2B6 efavirenz metaboliser phenotype distributions across world populations. The CYP2B6 phenotypes were predicted from diplotypes (see Table 2.2 for translation system) curated by PharmGKB researchers from various studies (`https://www.pharmgkb.org/page/cyp2b6RefMaterials`). The frequencies vary across populations with regards to the different categories: PM, poor metabolisers; IM, intermediate metabolisers; NM, normal metabolisers; RM, rapid metabolisers; UM, ultrarapid metabolisers. African populations have a relatively higher frequency of likely efavirenz poor metabolisers compared to European and Asian populations. This is largely due to the frequent *CYP2B6*6* and **18* alleles in African populations. Notably, even higher frequencies of the PM phenotype (~40%) have been reported in Oceanian populations (Mehlotra et al., 2006) — not shown in this graph due to the much lower sample size compared to other populations.

2.2.4.3 *CYP2B6* genotype to phenotype translation

Unlike *CYP2D6*, which has been extensively studied and for which several CPIC guidelines have been developed, there is only one CPIC guideline for a *CYP2B6* drug-gene pair i.e. *CYP2B6* and efavirenz (Desta et al., 2019). This guideline collates relevant published evidence and it provides dosing recommendations for efavirenz based on the *CYP2B6* genotype. Currently, the translation of *CYP2B6* genotype to the most likely efavirenz metaboliser status follows the recommendations summarised in Table 2.2.

Table 2.2: Translation of CYP2B6 genotypes to likely phenotypes in the context of efavirenz response.

Likely CYP2B6 phenotype	Genotype	Examples of *CYP2B6* diplotypes
Ultrarapid metaboliser	An individual carrying two increased function alleles	**4/*4, *22/*22, *4/*22*
Rapid metaboliser	An individual carrying one normal function allele and one increased function allele	**1/*4, *1/*22*
Normal metaboliser	An individual carrying two normal function alleles	**1/*1*
Intermediate metaboliser	An individual carrying one normal function allele and one decreased function allele OR one normal function allele and one no-function allele OR one increased function allele and one decreased function allele OR one increased function allele and one no-function allele	**1/*6, *1/*18, *4/*6, *4/*18, *6/*22, *18/*22*
Poor metaboliser	An individual carrying two decreased function alleles OR two no-function alleles OR one decreased function allele and one no-function allele	**6/*6, *18/*18, *6/*18*

This summary is adapted from (Desta et al., 2019)

As highlighted for *CYP2D6*, there are other factors that could influence *CYP2B6* activity aside from or in an interplay with an individual's *CYP2B6* genotype. These include drug-drug interactions involving *CYP2B6* inducers as well as enzyme inhibitors such as voriconazole and clopidogrel, and herbal DDIs which could all contribute to side effects or reduced efficacy to treatments such as efavirenz (Desta et al., 2016; Richter et al., 2004; Thomford et al., 2016). Furthermore, age and sex-related differences have been observed in *CYP2B6* expression for various substrates (Croom et al., 2009). For instance, Saitoh et al. (2007) observed a ~47% higher efavirenz clearance in children aged less than five years than in their older counterparts. These caveats underscore the need for more clinical and functional studies to go alongside variant discovery efforts in pharmacogenomics.

2.2.5 *CYP2A6* pharmacogenomic variation

The *CYP2A6* enzyme is involved in the metabolism of up to 3% of drugs that undergo biotransformation via CYP450 enzymes (Di et al., 2009; Shimada et al., 1994).

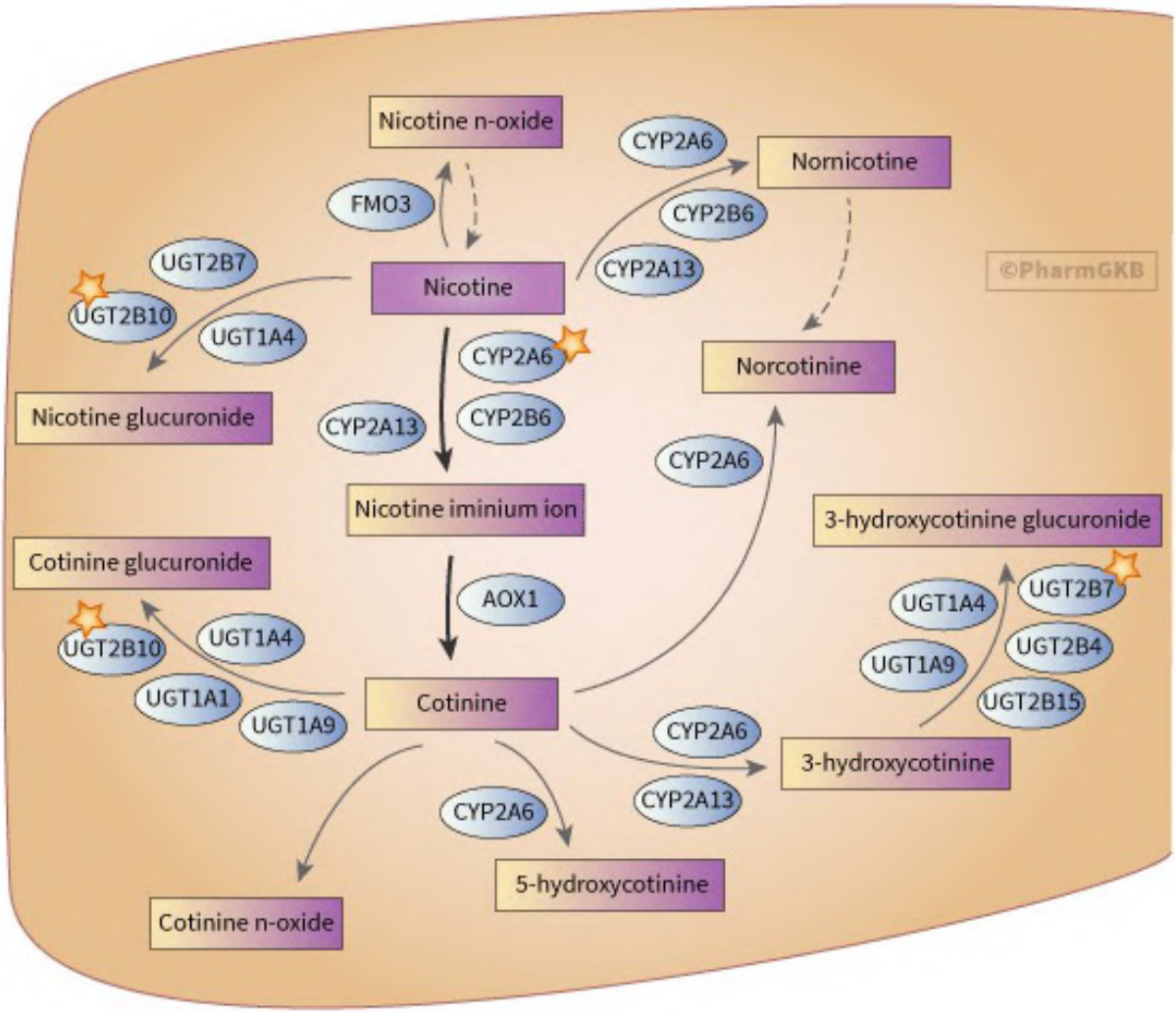


Figure 2.8: Schematic representation of the candidate genes involved in nicotine metabolism. CYP2A6 is the main CYP involved in the conversion of nicotine (~70-80%) cotinine (major pathway) in humans. The high *CYP2A6* polymorphism is known to cause inter-individual variability in this response. This image is reproduced from PharmGKB (`https://www.pharmgkb.org/pathway/PA2011`; see also Whirl-Carrillo et al., 2012) and is licensed under a Creative Commons BY-SA 4.0 license. AOX1, aldehyde oxidase; CYP, cytochrome P450; FMO3, flavin-containing monooxygenase 3; UGT, UDP-glucuronosyltransferase

---

CYP2A6 genotyping has contributed in particular to the understanding of interindividual variability in the metabolism of nicotine and other substrates, such as caffeine, antiretroviral agents (e.g. efavirenz) and anticancer agents (tegafur and letrozole) (Tanner and Tyndale, 2017; see also `https://www.pharmgkb.org/vip/PA166169430`). *CYP2A6* genetic variations have also been shown to influence smoking behaviors and could contribute to the risk for tobacco-

associated cancers (Strasser et al., 2011). The key involvement of *CYP2A6* in the nicotine pathway is highlighted in Figure 2.8.

2.2.5.1 Genomic architecture of the *CYP2A6* region

The *CYP2A6* gene is located on chromosome 19 (19q13) near the nonfunctional pseudogene *CYP2A7* (see Figure 2.9) within the large *CYP2ABFGST* gene cluster (Hoffman et al., 2001). *CYP2A6* and *CYP2A7* share a 95% sequence homology (Wassenaar et al., 2016) and there is also considerable similarity with *CYP2A13* (Hoffman et al., 2001). It is therefore no surprise that the *CYP2A* region harbours a number of structural variations (virtually analogous to *CYP2B6* and *CYP2D6*) (Oscarson et al., 1999; Rao et al., 2000). Importantly, these presence of the neighbouring *CYP2A7* and the structural variations makes amplification, sequencing and accurate bioinformatics *CYP2A6* variant analysis very challenging. This means that laborious methods, such as allele-specific PCR and Sanger sequencing, are often required in order to obtain reliable *CYP2A6* genotype calls (Wassenaar et al., 2016).

2.2.5.2 *CYP2A6* allele and phenotype diversity

As reviewed by Tanner and Tyndale (2017), recent GWAS studies of the nicotine metabolic ratio (NMR) and targeted resequencing efforts have provided important insights into the impact of *CYP2A6* genetic variation and differences in *CYP2A6* enzyme activity. To date, 45 *CYP2A6* star alleles have been identified but the full transition of the *CYP2A6* allele nomenclature from the preceding Human Cytochrome P450 Allele Nomenclature Database to PharmVar is still a work in progress (`https://www.pharmvar.org/gene/CYP2A6`). The frequencies of these alleles varies considerably within and between various populations and their activity is often characterised mainly with respect to nicotine metabolism (Nakajima et al., 2006; Piliguian et al., 2014). Figure 2.9b highlights some of the functionally significant allele-defining SNVs. *CYP2A6*23* (R203C), **24* (V110L and N438Y), **25* (F118L) and **28* (N418D and E419D), which have been identified in individuals of African ancestry are associated with a moderate decrease in the *CYP2A6* enzyme activity (Mwenifumbo et al., 2008; Tanner and Tyndale, 2017). Other alleles that cause a similar impact include *CYP2A6*7* allele (found predominantly in Asian populations), **5* (G479V), **18* (Y392F), and **21* (K476R). *CYP2A6*17* (V365M) and *CYP2A6*20* (2141_2142delAA, frameshift) are both African-specific (AF of 7–11% and 1–1.7%, respectively) and they cause a considerable decrease or complete loss of activity towards *CYP2A6*-mediated nicotine metabolism (Fukami et al., 2004; Tanner and Tyndale, 2017), similar to *CYP2A6*2* (L160H) and **10* (I471T and R485L), which are predominant in Caucasian and Asian populations, respectively (Tanner and Tyndale, 2017; Yoshida et al., 2002; Y. Zhou et al., 2017).

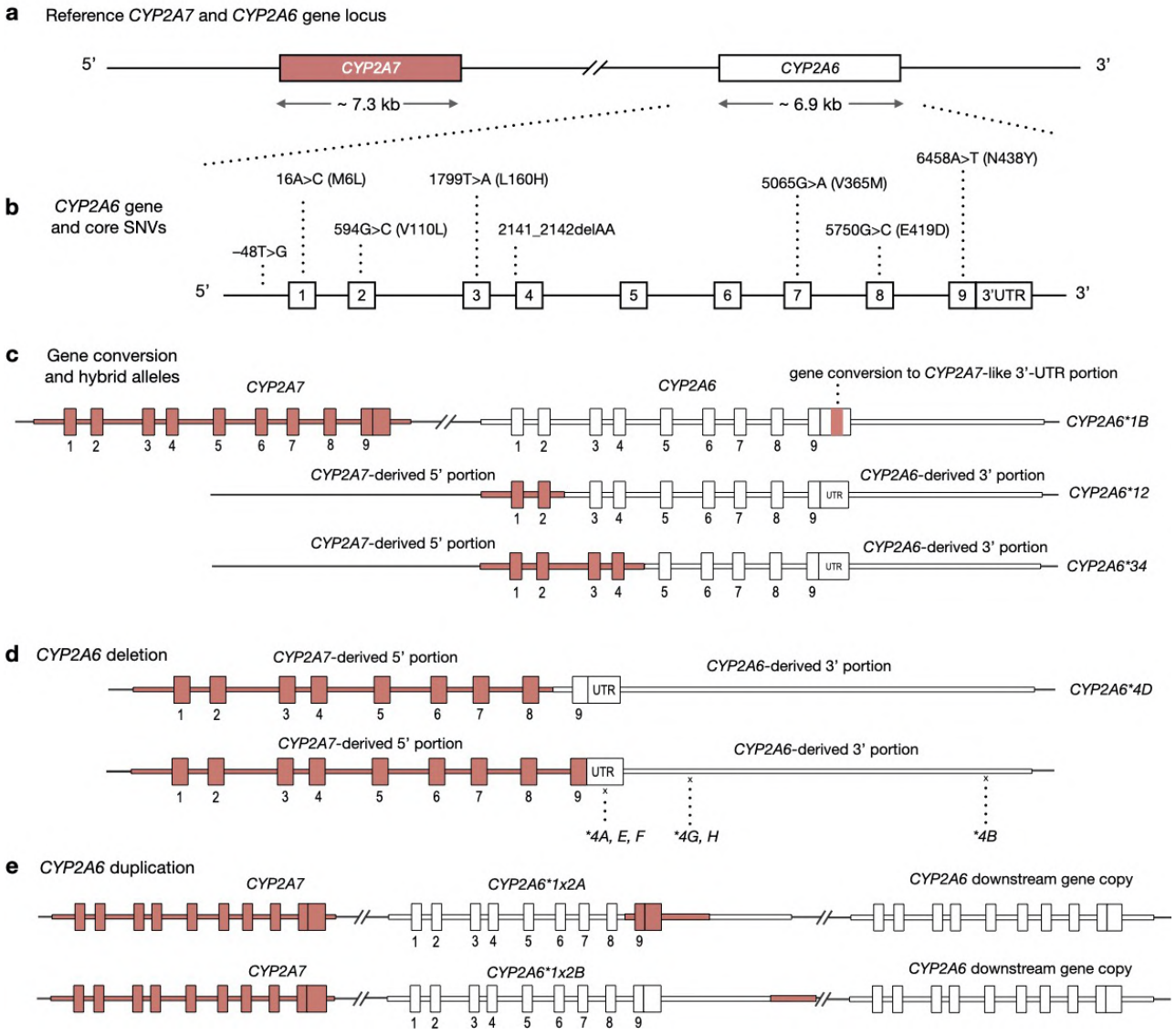


Figure 2.9: Graphical overview of the *CYP2A6* gene locus and allelic variation. (**a**) Shows the *CYP2A6* gene on chromosome 19 neighbouring the highly similar but nonfunctional pseudogene, *CYP2A7*. The locus is shown here in reverse since it occurs on the minus strand. (**b**) Schematic showing the location of selected *CYP2A6* SNVs — with a focus on those identified in African populations (discussed in main text) — relative to the ATG-start on the RefSeqGene (NG_008377.1). (**c**) Visualisation of the common *CYP2A6* gene conversion to *CYP2A7* in the 3' UTR (*CYP2A6*1B*), and *CYP2A7-2A6* fusion products (**12* and **34*) that are thought to have resulted from unequal recombination. (d) Schematic representation of the *CYP2A6* deletion hybrid allele (**4*, nonfunctional), including the various breakpoints that have been characterised so far in **4A*, **4B* and **4D-H*. (e) Depiction of *CYP2A6* duplications — likely reciprocal products to the gene deletion during unequal crossover. The duplicated copies have *CYP2A7*-derived 3'-ends. The switch region in CYP2A6*1x2A is in the intron 8 - 3'UTR region (Rao et al., 2000), while the switch region in CYP2A6*1x2B is ~5.2-5.6 kb downstream from the stop codon (Fukami et al., 2007). A more detailed catalogue of these alleles is available at `https://www.pharmvar.org/gene/CYP2A6`.

Regarding the aforementioned *CYP2A6* structural variations, these include gene duplications, deletion, conversions, and hybrid alleles (Figure 2.9 c-e), which are believed to be due to unequal crossover recombination events between *CYP2A6* and *CYP2A7* (Mwenifumbo et al., 2010; Rao et al., 2000). The *CYP2A6*1B* allele, which is characterized by a 58-bp gene conversion in the 3' region (Ariyoshi et al., 2000) (see Figure 2.9c), is associated with increased *CYP2A6* mRNA stability and thus higher protein levels (J. Wang et al., 2006). However, the *CYP2A6*1B* activity has not been fully ascertained as it is often found in LD with decreased function variants (Tanner and Tyndale, 2017). Furthermore, Al Koudsi et al. (2010) observed no differences in *CYP2A6* mRNA, protein, or activity between *CYP2A6*1A/*1A* and *CYP2A6*1A/*1B* or *CYP2A6*1B/*1B* genotype groups in a human liver bank.

There are two known gene duplication alleles to date i.e. *CYP2A6*1x2A* (Rao et al., 2000) and *CYP2A6*1x2B* (Fukami et al., 2004) (see Figure 2.9e), and these are both associated with faster nicotine metabolism. Conversely, *CYP2A6*4* (gene deletion) is nonfunctional (Oscarson et al., 1999). *CYP2A6*4* occurs at a frequency of 0.5–2.7% in individuals of African ancestry and it is particularly frequent in Alaska natives (AF up to 15%) (Tanner and Tyndale, 2017). There are various *CYP2A6*4* suballeles (e.g. **4A*, **4B*, **4E* and **4G*) due to differences in the breakpoint positions (see Figure 2.9d) (`https://www.pharmvar.org/gene/CYP2A6`).

In addition to the CNVs, two *CYP2A7-2A6* hybrids (*CYP2A6*12* and *CYP2A6*34*) have also been characterised so far (di Lulio et al., 2009; Oscarson et al., 2002). Exons 1-2 of *CYP2A6*12* are derived from *CYP2A7* while exons 3-9 are of *CYP2A6* origin. *CYP2A6*34* has *CYP2A7*-like exons 1-4 while exons 5-9 are derived from *CYP2A6*. While *CYP2A6*12* is known to be associated with decreased *CYP2A6* enzyme activity, the activity of *CYP2A6*34* is unknown.

Generally, African American and Asian populations tend to have higher frequencies of *CYP2A6* decreased function or nonfunctional alleles, hence exhibiting lower *CYP2A6* activity compared to individuals of European ancestry (Tanner and Tyndale, 2017). However, more studies are needed to get a better understanding of the *CYP2A6* allele distributions and relevant phenotypes in continental African populations.

2.2.5.3 Other factors influencing CYP2A6 metabolic activity

It is important to note that the activity of *CYP2A6* can be influenced by factors other than the allelic variation.

From the induction perspective, *CYP2A6* expression and activity may be induced by drugs, endogenous substances and dietary components (reviewed by Tanner and Tyndale, 2017). For example, drugs such as phenobarbital, dexamethasone, and rifampin could activate nuclear hormone receptors (e.g. the pregnane X receptor, PXR) leading to increased *CYP2A6* transcription (Itoh et al., 2006; Rae et al., 2001). Similarly, there is a growing body of evidence suggesting

that estrogen is associated with greater *CYP2A6* activity among the endogenous transcription inducers (Higashi et al., 2007). Among the dietary substances, broccoli has been associated with raising levels of *CYP2A6* activity when consumed in large quantities (500g over six days) (Hakooz and Hamdan, 2007).

From the inhibition point of view, drugs, endogenous substances and dietary components may act as inhibitors to the *CYP2A6* enzyme. Drugs such as methoxsalen (treats skin conditions) and selegiline (used in early Parkinson's disease treatement) are bioactivated by *CYP2A6*, and thereafter their products irreversibly bind to and inhibit *CYP2A6* (mechanism-based inhibition) (Koenigs et al., 1997; Siu and Tyndale, 2008). *CYP2A6* inhibition by endogenous substances has been demonstrated for dopamine, serotonin and tryptamine (Higashi et al., 2007; W. Zhang et al., 2001). From the dietary perspective, grapefruit juice, menthol (common flavourant), and coffee constituents (e.g. caffeic acid, p-coumaric acid, and quercetin) have been shown to inhibit *CYP2A6* as well (Hukkanen et al., 2006; MacDougall et al., 2003; Ware et al., 2017. In essence, lifestyle, age, sex and physiological factors need to be considered alongside genetic variation while developing and implementing clinical pharmacogenomics guidelines.

2.3 Catalysing pharmacogenomics and precision medicine in Africa: opportunities and challenges

2.3.1 Next-generation sequencing technologies

The more widely adopted first generation sequencing technology, namely Sanger sequencing (Sanger et al., 1977), was central to producing the draft sequence of the human genome, which was published two decades ago (International Human Genome Sequencing Consortium, 2001). Sanger sequencing is a sequencing-by-synthesis (SBS) approach i.e. it requires direct action by the enzyme DNA polymerase to extend a DNA primer, which is essentially a starting point bound next to the region of interest (Morozova and Marra, 2008; Sanger et al., 1977). The key feature of Sanger sequencing is the use of chain-terminating nucleotides called dideoxynucleotide triphosphates (ddNTPs) alongside normal (A, G, T, C) building blocks in four individual DNA synthesis reactions (Sanger et al., 1977). The ddNTPs cause DNA synthesis to be terminated at many different positions for each reaction. The order of the nucleotides in the DNA sequence is determined after separating the products of the synthesis reactions using gel (or capillary) electrophoresis, a method that separates DNA fragments by their sizes (Applied Biosystems, 2009).

Sanger sequencing produces highly accurate (99.99%) DNA reads averaging 500-800 bp in length (Morozova and Marra, 2008; Sanger et al., 1977) and is often used in combination with polymerase chain reaction (PCR), which facilitates DNA amplification for downstream applications (Saiki et al., 1988). In the pharmacogenetics perspective, Sanger sequencing has

been central to the characterisation of important pharmacogene regions such as the complex *CYP2D6* gene locus (Kimura et al., 1989), and remains a gold standard for the validation of novel pharmacogene variants (Desta et al., 2021; Gaedigk, Turner, et al., 2019; Nofziger et al., 2020). However, Sanger sequencing's low throughput, high DNA template requirement and laborious protocols make it expensive and unsuitable for investigating genome-wide variation across populations (Heather and Chain, 2016).

As in most genomics fields, pharmacogenomics has greatly benefited from the emergence of next generation sequencing technologies, which have provided the potential for large-scale investigations of virtually all types of genomic variations (Koboldt et al., 2013; Wright et al., 2018; S.-F. Zhou, 2009). Next Generation Sequencing (NGS) technologies are characterised by the ability to sequence numerous DNA (or cDNA) fragments in parallel. In this way, NGS technologies offer high sequence yield, greater read depth, much shorter turnaround times ($\leq$ one week for human whole genome sequencing (WGS) on average), and remarkable cost-effectiveness compared to Sanger sequencing (Heather and Chain, 2016). Two of the main categories used to classify NGS technologies include the 'second-generation' and 'third-generation' categories, the key difference being that the latter are capable of sequencing single DNA molecules (Heather and Chain, 2016; McGinn and Gut, 2013).

2.3.1.1 Second-generation technologies

Among the second-generation sequencing advancements, the SBS approach developed by Solexa (later acquired by Illumina; `https://www.illumina.com/`) has arguably been the largest contributor and driver for genomics research since the completion of the Human Genome Project (D. R. Bentley et al., 2008; Goodwin et al., 2016). Illumina has four major sequencing platforms namely the NovaSeq, HiSeq, NextSeq, and MiSeq instruments (`https://www.illumina.com/`). Competing platforms to Illumina include the sequencing by oligonucleotide ligation and detection (SOLiD) system from Applied Biosystems (McKernan et al., 2009); the Ion Torrent — a semiconductor platform (Rothberg et al., 2011); and the MGISEQ and BGISEQ models developed by the Beijing Genomics Institute (BGI) Group (Huang et al., 2017; Jeon et al., 2019). These second-generation sequencing platforms generate relatively short DNA read lengths averaging 100bp (range: 75bp - 600bp) (Zook et al., 2019), with per base call accuracy >99% (up to 99.9% for Illumina sequencing). Unlike the Ion Torrent platform, Illumina, SOLiD and BGI platforms have the capability of producing paired-end reads for each DNA segment which improves the accuracy when aligning reads to reference sequences, especially across regions with repetitive sequences (Heather and Chain, 2016).

Despite the remarkable accuracy and cost-effectiveness of second-generation technologies, they struggle with reading genomic regions with homopolymers and repetitive sequences (Heather

and Chain, 2016). Additionally, the short DNA reads they generate are not ideal for variant phasing across long stretches in the genome (see Figure 2.10b), unless pedigrees (genetic data from a family tree) are used. Furthermore, structural variant detection based on short-read data is often inconsistent especially in gene loci (e.g. the *CYP2D6* region) with highly homologous pseudogenes which cause read mapping ambiguities, especially in the presence of complex hybrid gene rearrangements (Desta et al., 2021; Twist et al., 2016; Yang et al., 2017) (see Figure 2.10).

In spite of these short-comings, short-read sequencing with support from a wide range of analysis tools and pipelines has undoubtedly enhanced population-level pharmacogenomics research and clinically-relevant variant discovery globally (Goodwin et al., 2016; Wright et al., 2018; Y. Zhou et al., 2017).

2.3.1.2 Third-generation technologies

The emergence of third-generation NGS technologies (long-read sequencers) has created a new wave of excitment and opportunity in genomics research. These sequencers are capable of reading single molecules, thus negating the need for DNA amplification and promoting generation of long sequence reads given that the DNA quality is robust. Currently, the popular third-generation NGS technologies include the nanopore sequencing platforms from the Oxford Nanopore Technologies (ONT) (`https://nanoporetech.com`) and the single molecule real-time (SMRT) sequencing platforms from Pacific Biosciences (`https://www.pacb.com`). These are henceforth referred to as Nanopore and SMRT sequencing technologies, respectively.

The most commonly used Nanopore platforms are the GridION, the mobile phone-sized MinION, and the PromethION (Liau et al., 2019; Zook et al., 2016). These sequencers measure the ionic current disruptions when single-stranded nucleic acids pass through nanopores embedded in an electro-resistant biological membrane (`https://nanoporetech.com`). Nanopore sequencers produce long reads (up to >2Mb) (Amarasinghe et al., 2020; Payne et al., 2019). Their read length is essentially determined by the length of input DNA molecules.

SMRT sequencers (e.g. Sequel II, and Sequel IIe) detect fluorescently tagged nucleotides (via real-time imaging) as they are being synthesised along individual DNA template molecules by a DNA polymerase (Roberts et al., 2013; see also `https://www.pacb.com`). SMRT sequencers produce 10-15 kb reads on average, depending on the application, and they can sequence up to 30kb DNA fragments (Amarasinghe et al., 2020).

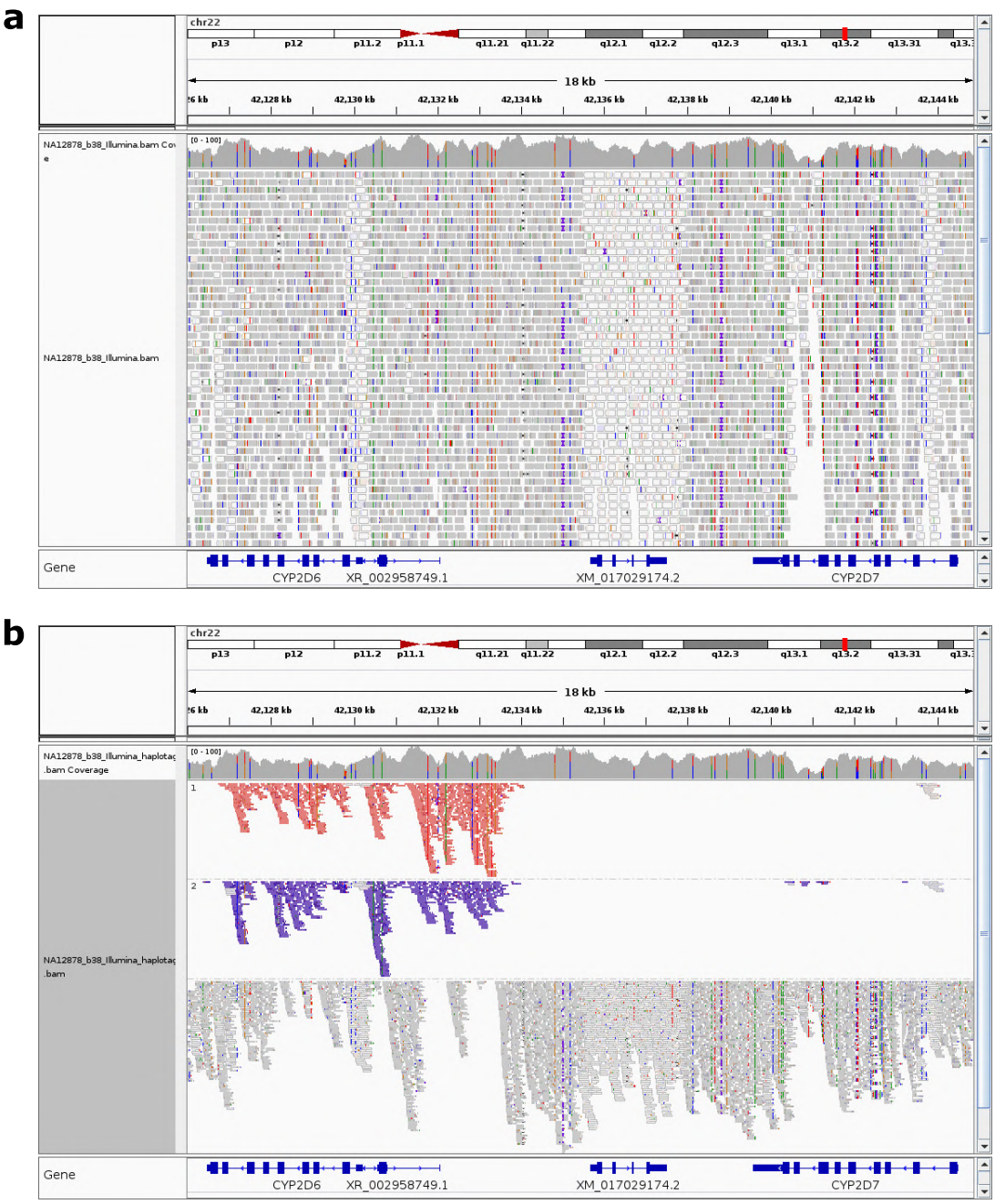


Figure 2.10: Visualisation of the Illumina paired-end read alignment to GRCh38 for sample NA12878 (*CYP2D6*3/*68+*4* diplotype) — WGS data generated by Zook et al. (2016). (**a**) Shows aligned reads before haplotagging i.e. colouring reads resolved to be from one haplotype differently from those resolved to be from the second haplotype. The **68* allele is a *CYP2D6-2D7* hybrid as reviewed earlier. The short-reads from the *CYP2D7*-derived 3' portion of the **68* hybrid are actually aligned to *CYP2D7* hence the jumbled coverage seen in the figure. Other non-specific read alignments due to the high homology between *CYP2D6* and *CYP2D7* are also a factor. (**b**) Depicts the haplotagged reads resolved using Whatshap (Patterson et al., 2015). The reads in red are mainly from the *3 haplotype while the reads in purple are mainly from the *4 haplotype. However, it is difficult to unequivocally phase all the variants from the **3* and **4* alleles and to correctly call the **68+*4* tandem arrangement using the read data alone. Furthermore, there are numerous reads (gray-coloured) that cannot be used for SNV phasing purposes as they do not harbour at least two sequence variants. Statistical phasing, pedigrees and/or combinatorial approaches are therefore the go-to stategies to infer *CYP2D6* alleles from short-read data. This image was produced using the Integrative Genomics Viewer (Robinson et al., 2011).

Due to the ability to produce long reads, Nanopore and SMRT sequencers provide the opportunity for comprehensive genome assemblies (Aganezov et al., 2021; Jain et al., 2018; Miga et al., 2020), haplotype validation (Buermans et al., 2017, Liau et al., 2019; Qiao et al., 2016), transcriptomics, characterising structural variations (Jiang et al., 2020), and studying the methylome (Liu et al., 2020). In addition, the Nanopore platforms in particular offer other advantages including portability and real-time sequence data analysis, making them easy to set up in laboratories and field operations (Bull et al., 2020). In the pharmacogenetics perspective, the haplotyping (see Figure 2.11) and enhanced structural variant characterisation advantages are appealing especially for interrogating the complex *CYP2D6* locus as well as the large *CYP2ABFGST* gene cluster. For example, Liau et al. (2019) validated the application of Nanopore sequencing to *CYP2D6* haplotyping and structural variant detection, while Buermans et al. (2017) and Qiao et al. (2016) validated SMRT sequencing protocols for various *CYP2D6* allelic arrangements. The targeted SMRT-sequencing-based allele characterisation by Qiao et al. (2016) led to the definition of *CYP2D6*107* through **109*, while the Nanopore-based haplotype validation by Liau et al. (2019) led to the characterisation of *CYP2D6*127* and multiple suballeles.

Despite their clear advantages over short-read technologies, long-read sequencers are yet to attain the same level of ubiquity across clinical and research settings. This is in part due to the relatively lower throughput and higher cost of sequencing associated with both Nanopore and SMRT sequencing compared to Illumina and Ion Torrent platforms for the same genome coverage or read depth (Amarasinghe et al., 2020; Das et al., 2019). Furthermore, long-read sequencing platforms have historically been associated with a relatively high per-base sequence error rate compared to short-read sequencing platforms (Das et al., 2019; Koren et al., 2012). For example, Dohm et al. (2020) found a high rate of insertions (~8%) erroneously introduced during their SMRT sequencing run whereas Nanopore reads showed approximately a 4% error-rate for substitutions, insertions and deletions.

Recent versions of the Nanopore and SMRT sequencing platforms are being developed with increased mitigation of these sequencing errors as a major target (Rang et al., 2018). For example, the MinION has undergone iterations from R6 chemistry (~60% accuracy) to R10 (accuracy over 90%) for 1D read sequencing i.e. passing a single DNA strand through the nanopore (`https://nanoporetech.com`). Other Nanopore enhancements include supporting $1D^2$ (~97% accuracy) sequencing, which enables sequencing of both the template and complementary DNA strand without physical ligation (Rang et al., 2018). Remarkably, SMRT sequencing is currently able to achieve long (10-15kb on average) reads with an accuracy of over 99.5% consistently by reading the same molecule multiple times (~10 passes on average – generating a subread each time) and then obtaining consensus (CCS) high fidelity (HiFi) reads (Wenger et al., 2019).

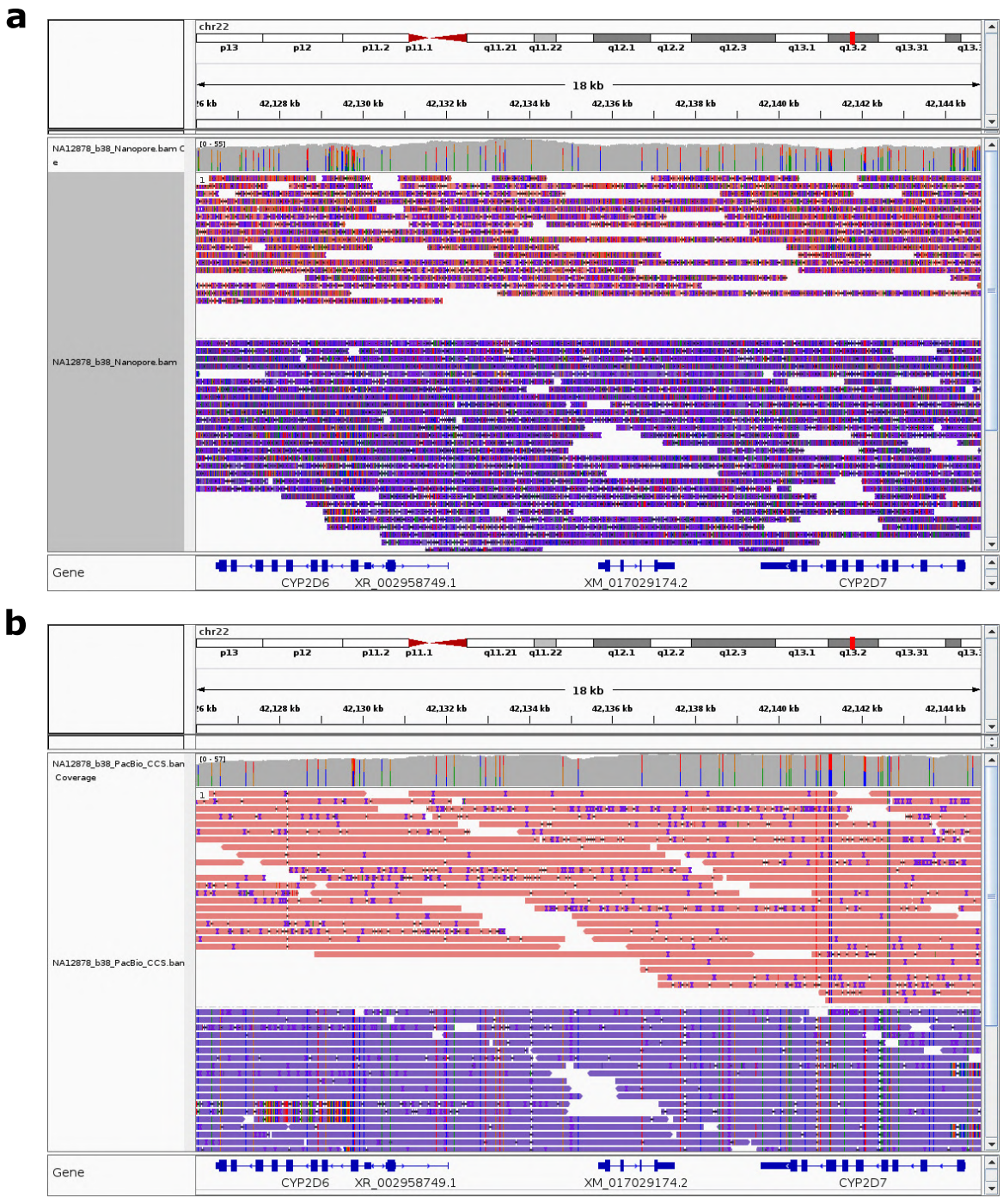


Figure 2.11: Visualisation of the long-read alignment in the *CYP2D* region (GRCh38) for sample NA12878 (*CYP2D6*3/*68+*4* diplotype) – WGS data generated by Zook et al. (2016). (**a**) Depiction of the haplotagged Nanopore read alignment for the *CYP2D* region in NA12878. SNVs from the two haplotypes i.e. **3* (red-coloured reads) and **4* (purple-coloured reads) can easily be phased. However, the individual reads have high error-rate (see purple spots in the red-coloured reads showing false-positive insertions). Notably, the reads from the **68 CYP2D6-2D7* hybrid are mapped ambiguously. (**b**) Visualisation of the haplotagged PacBio read alignment for the *CYP2D* region in NA12878. The PacBio reads generated from circular consensus sequencing (CCS) have much fewer errors compared to the Nanopore reads in (a). Core SNVs (and subvariants) defining **3* and **4* can be phased easily since the long reads span the entire *CYP2D6* gene, however the reads from the **68* hybrid are aligned to *CYP2D7* since the switch region occurs towards the beginning of the gene (intron 1).

Some of the aforementioned next-generation technologies' shortcomings to date are being offset by parallel advances in NGS data analysis tools and pipelines. These include hybrid error-correction algorithms for analysis of long-read sequencing data (Dohm et al., 2020; LoTempio et al., 2021), which often leverage the accuracy of short-read data from second-generation sequencers and the read length from third-generation sequencers.

2.3.1.3 Array-based technologies and panel-based assays

Even though this literature review has so far been focused on the more dominant next generation technologies, there are other technologies that either compete with or complement NGS platforms in various genomic applications. These alternative technologies are mainly not sequencing-based i.e. they are used to genotype common variants or a few variants of interest rather than surveying all nucleotides in the genome.

Microarrays are perhaps the most widely used alternative to mainstream NGS sequencers. DNA microarrays (genotyping chips) are used to identify common (and often medically-relevant) SNPs in a population at a much cheaper cost compared to NGS routines. Thus, DNA microarrays have facilitated applications such as genome wide association studies (GWAS) which investigate SNP associations to phenotypic traits (Diouf et al., 2015; Hayat et al., 2020), including drug response (Ovenden et al., 2017). Microarrays are also often the platform of choice for studies on ethnicity, population structure and human migration (Sengupta et al., 2021; Tishkoff et al., 2009) all of which could influence the distribution of various pharmacogenomic variations. Other types of microarrays include Comparative Genomic Hybridisation arrays (aCGH) which are used to identify structural variations in the genome (Roy and Motsinger-Reif, 2013), and expression arrays used to measure the expression levels of various genes in a cost-effective manner as an alternative to RNA sequencing (RNA-seq) (Liang et al., 2013).

As expected, the main drawback to using microarrays is that they are limited when it comes to assessing rare and/or novel variation. Ungenotyped rare variants in a study population can be inferred using information from densely typed SNPs but this would require comprehensive reference panel(s) for it to be accurate (Sariya et al., 2019; Vergara et al., 2018). The widely used reference panels currently include the 1000 Genomes Project reference panel (Auton et al., 2015), the Haplotype Reference Consortium reference panel (release 1) (McCarthy et al., 2016), and the UK10K reference panel (Huang et al., 2015). However, these panels predominantly comprise genomic sites from samples of European ancestry. This disparity is being addressed e.g. by recent development of the African Genome Resources haplotype reference panel (`https://imputation.sanger.ac.uk`), but the African ethnolinguistic groups being represented are few.

As microarrays are typically applied in genome-wide SNP genotyping, there are a number of similar platforms/assays available for targeted pharmacogenetics applications in CLIA certified

laboratories (Fernandez et al., 2012; Gaedigk, Turner, et al., 2019). These platforms are mainly based on array technology or NGS panel enrichment. NGS panel-based platforms are often used in combination with long-range (XL)/allele-specific PCR amplification and short-read sequencing. Table 2.3 provides a summary of some of the currently available alternative pharmacogenetics platforms. The strengths and limitations of these platforms differ by design. For instance, panel-based platforms such as the PGRNseq (Gordon et al., 2016) enable detection of novel alleles and a few structural variants in genes such as *CYP2D6* and *CYP2A6*, while array-based platforms such as DMET Plus (Affymetrix) and AmpliChip CYP450 Test (Roche; discontinued by the developers) can only genotype SNVs included on the array. Various combinations of these platforms are often used to enable genotyping of haplotypes defined by SNVs and/or structural variants in complex pharmacogenes such as *CYP2D6* (Gaedigk, Turner, et al., 2019). Notably, most of these assays have been designed based on allele data from populations of European ancestry and have been shown to miss rare and novel haplotypes in African populations (Dodgen et al., 2013).

2.3.1.4 Short-read long-insert technologies

Besides microarrays and panels, another notable NGS strategy is the use of 'short-read long-insert' technologies such as optical mapping, linked-read, and high-throughput chromosome conformation capture (Hi-C) technologies (Eisenstein, 2015). Optical mapping technology, provided predominantly by Bionano Genomics (`https://bionanogenomics.com`), involves combining long-read technology with low-resolution sequencing (opposite of base-by-base DNA sequencing). Fluorescent markers are used to tag specific motifs within DNA fragments that are up to ~1 Mb in length, and thereafter, images of the fluorescent signal patterns are then used to produce genome maps via alignment to each other and/or a reference sequence (Goodwin et al., 2016; Yuan et al., 2020). Optical mapping facilitates gap filling and scaffolding in short-read-based *de novo* sequence assemblies and it also aids in studying structural variations as well as haplotype blocks which span up to tens or hundreds of thousand kilobases in length, respectively (Cao et al., 2014; Shelton et al., 2015). Optical mapping is effective in characterising structural variations in intractable parts of the genome i.e. repetitive regions, segmental duplications and telomeric areas (Levy-Sakin et al., 2019), but its relatively high error rate, lack of base-pair resolution, and the need for expensive specialised technology platforms are major deterrents to its routine use (Ho et al., 2020; Yuan et al., 2020).

The linked-read strategy (10X Genomics; `https://www.10xgenomics.com`) differs from optical mapping in that it entails production of 'synthetic long reads' via specific modifications to the short-read sequencing library preparation protocols. The 10X Genomics linked-read platform partitions cells or nuclei (using a microfluidic system) and prepares sequencing libraries in parallel, ensuring that all fragments (from diluted high-molecular-weight DNA) produced

within a partition share a common barcode – delivered by a gel bead. The fragments then undergo short-read sequencing from which long-range information is reconstructed (using *in silico* pipelines) by determining the origin of the short-read fragments from their respective barcodes (Tourdot et al., 2021; Zheng et al., 2016). Linked reads retain their underlying short-read information in terms of the accuracy and depth coverage for each fragment thus making them well-suited for structural variant detection (Ho et al., 2020). In addition, the combination of low error rate and long molecule length (up to 100 kb) make this approach suitable for haplotype phasing as illustrated in Figure 2.12.

In spite of these strengths, the utility of linked-read technology is limited as it is more expensive than routine short-read sequencing due to the preceding barcoding library preparation steps. Another main drawback of linked-read technology is the low detection rate for insertions in the genome due to coverage gaps within DNA fragments especially in repetitive regions (Tourdot et al., 2021).

In contrast to both optical mapping and linked-read technologies, Hi-C involves sequencing cross-linked chromatin (DNA and protein macromolecular complex in the nucleus) to analyse DNA sequences that may be distant in the linear genome but proximal in the 3D genome architecture (Harewood et al., 2017). Hi-C is useful for detecting large structural variations, especially translocations, as the read pairs it generates can span megabases. However, a deeper understanding of the genome spatial organisation is necessary before the full potential of the Hi-C technology is realised — especially given that regular fluctuations in the 3D chromatin structure present difficulties in identifying potential structural variations (Ho et al., 2020).

In a nutshell, no single NGS technology is comprehensive enough to enable detection of all kinds of genomic variations. Additionally, there is room for improvement regarding the time and cost of sequencing especially for long-read sequencing (Sedlazeck et al., 2018). In the same vein, despite the plethora of bioinformatics tools and pipelines in research and clinical settings, workarounds to some of the technological limitations may or may not be feasible. Hybrid approaches involving the different sequencing strategies as well as analysis pipelines could be the catalyst for a more complete characterisation of SNVs and structural variations in various populations, and for providing benchmarks (Zook et al., 2019). However, this would require a bigger drop in the cost of sequencing to be feasible in understudied populations. Nonetheless, with their high throughput and relatively cheaper data generation, next generation technologies provide the opportunity to expand our understanding of population-specific variation in under-represented regions (e.g. in Africa) across the genome, and in particular for pharmacogenes.

Table 2.3: Pharmacogene genotyping platforms that rely on array, PCR and/or NGS target enrichment assays.

Manufacturer/ Designers	Product/Assay	Genes investigated	Variants/panel size	Technology
Applied Biosystems	TaqMan DME SNV assays	∼219 ADME genes	Over 3000 protein-coding SNVs	Real-Time PCR assay
Applied Biosystems	TaqMan CNV assay	*CYP2A6*, *CYP2D6*, *GSTM1*, *GSTT1*, *SULT1A1*, *UGT2B17*	Gene deletion and duplication events (unphased)	Real-Time PCR assay
Thermo Fisher Scientific	PharmacoScan	∼1191 pharmacogenes (including ADME and drug target genes)	∼4,627 markers CN (0, 1, 2, ≥3)	Microarray; Multiplex PCR for loci with pseudogenes
Thermo Fisher Scientific	TaqMan OpenArray PGx Express Panel	15 pharmacogenes (including *CYP2D6*)	60 SNVs	TaqMan genotyping
Agena Bioscience	VeriDose Core Panel	20 ADME genes (e.g. *CYP2D6* and *CYP2C19*)	68 SNVs	Single-base extension; MassARRAY system
Agena Bioscience	VeriDose CNV panel	*CYP2D6*	6 *CYP2D6* CNV regions	CNV panel; MassARRAY system
Custom (CMKC) (Gaedigk et al., 2019)	MPA CNV	*CYP2D6*	4 *CYP2D6* CNV regions	Quantitative multiplex PCR
Roche Molecular Systems	AmpliChip CYP450 Test (discontinued)	*CYP2C19*, *CYP2D6*	33 *CYP2D6* alleles; 3 CYP2C19 alleles; *CYP2D6* CNVs	GeneChip microarray
Affymetrix	DMET™ Plus	∼225 ADME genes	∼1,931 variants; 5 CNV regions	GeneChip microarray
Illumina	VeraCode®	ADME core genes panel	184 biomarkers in 34 genes	Beads microarray
Illumina	GDA+PGx	2000 targets (including *CYP2D6*, *CYP2B6* and *TPMT*)	44,000 PGx markers genome-wide	Microarray; TGA for *CYP2D6*, *CYP2B6* and *TPMT*
Custom (Gordon et al., 2016)	PGRNseq	∼84 pharmacogenes	∼1325 SNVs per individual	Target enrichment; Illumina sequencing
Custom (Gulilat et al., 2019)	PGx-seq	∼100 pharmacogenes	Over 1000 SNVs	Target enrichment (exome only); Illumina sequencing
Custom (CMKC)	ADMEseq	Over 280 ADME genes	Panel size of ∼660 kb	Target enrichment; Illumina sequencing
Custom (K. Klein et al., 2019)	ADME NGS panel	340 ADME genes	Panel size of ∼1,382 kb	Target enrichment; Illumina sequencing
Bio-Rad	ddPCR	Multiple	Multiple CNV regions	Water-oil emulsion droplet technology
Luminex	xTAG *CYP2D6* Kit v3	*CYP2D6*	Multiple *CYP2D6* SNVs	Multiplex PCR; Allele-specific primer extension

CN, copy number; ddPCR, digital droplet PCR; MPA, Multiplex Probe Amplification CMKC, Children's Mercy Hospital, Kansas City; TGA, Targeted Gene Amplification; GDA+PGx, Infinium Global Diversity Array with Enhanced PGx.

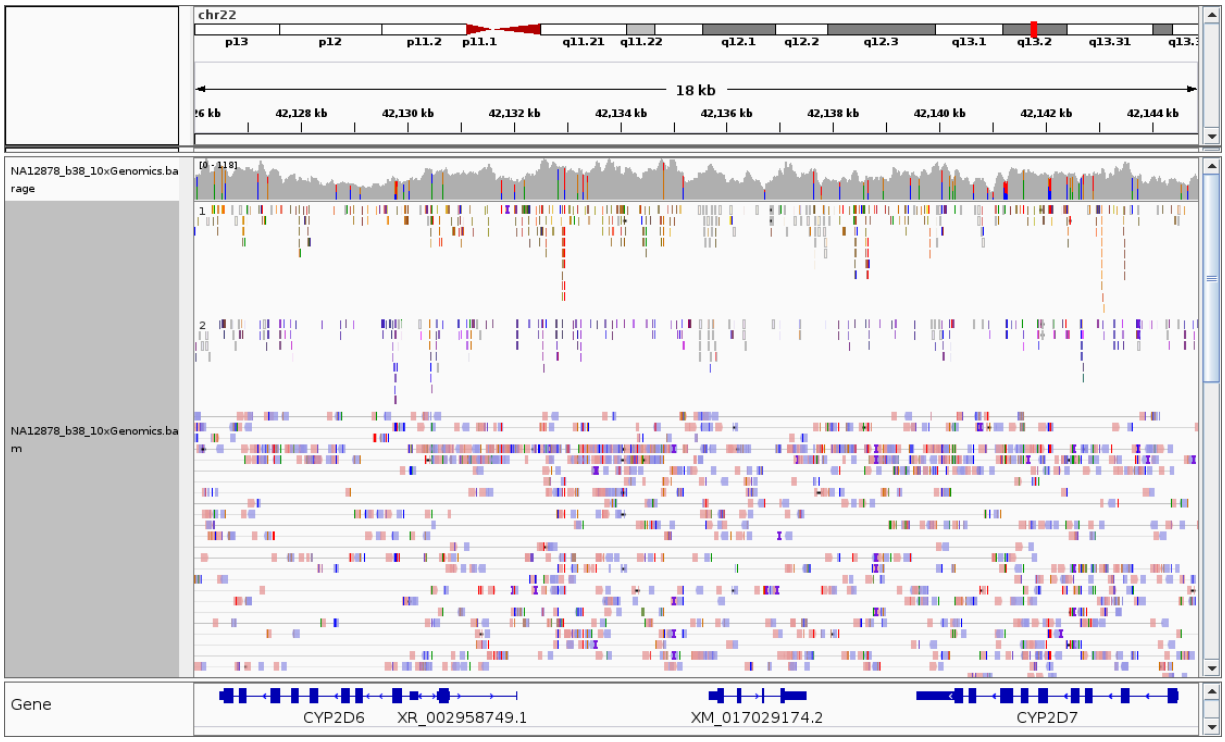


Figure 2.12: Visualisation of the 10X Genomics read alignment to GRCh38 (*CYP2D* region) for sample NA12878 (*CYP2D6*3/*68+*4* diplotype) — WGS data generated by Zook et al. (2016). The 10X Genomics sequencing technology uses barcoding strategies so as to generate 'synthetic long reads' that leverage the accuracy of short reads. The barcoded reads — originally from the same DNA fragment — can be inferred bioinformatically (see reads joined together by the gray lines in the figure), thus facilitating phasing of the SNVs harboured by these reads. It is however difficult to use this technology to interrogate structural variations due to the gaps in the genome read coverage.

2.3.2 Emergence of star allele calling bioinformatics pipelines

Given the enhanced sequencing ability afforded by next generation technologies, bioinformatics tools for automated pharmacogene variant detection, haplotype inference and/or phenotype prediction have become increasingly important in clinical and research settings (Twist et al., 2016; Dalton et al., 2020).

2.3.2.1 Existing star allele callers

The tools/pipelines that have recently been developed for pharmacogene star allele calling using high-throughput sequence data include Cypiripi (now deprecated by its developers) (Numanagić et al., 2015), Astrolabe (Twist et al., 2016), Aldy (Numanagić et al., 2018), Stargazer (S.-B. Lee, Wheeler, Patterson, et al., 2019), PharmCAT (Sangkuhl et al., 2020), PGxPOP (McInnes et al., 2021), and Cyrius (Chen et al., 2021). These tools typically have to analyse short-read sequence data since it is more ubiquitous than long-read sequence data as described in section 2.3.1. However, some tools such as PharmCAT and PGxPOP, which use only variant call format

(VCF) files as input, do not have algorithms for calling structural variants at present. Given the complex architecture for genes such as *CYP2D6*, *CYP2A6* and *CYP2B6* (section 2.2), extra caution has to be taken while interpreting allele calls from all the *in silico* tools for such highly polymorphic genes.

So far, star allele calling bioinformatics algorithms have proved to be useful in population-scale HTS pharmacogenomics analyses by enabling baseline allele and phenotype frequency estimations, and narrowing down the number of samples requiring orthogonal haplotype validation. For example, (Mauleekoonphairoj et al., 2020) used Stargazer for automated star allele calling and phenotype prediction for 25 pharmacogenes in a study comprising 291 Thai individuals within a Brugada cohort. Another illustration of the efficiency rendered by such algorithms was the application of PGxPOP and PharmCAT for the star allele analysis using over 400,000 samples (VCF format) from the UK Biobank (McInnes et al., 2021).

Currently, the adoption of the available allele callers in the clinical setting is still hesitant in part because a comprehensive benchmark analysis is lacking (Nofziger et al., 2020). It is also important to note that these tools have not been extensively used to study pharmacogene variation in African populations, except for some samples included in publicly available datasets such as the 1000 Genomes Project datasets (Auton et al., 2015). Genotyping samples with novel alleles could be particularly challenging since these alleles would typically be missing from the databases used by the algorithms in diplotype assignment. It is unclear as to how the high genetic diversity in Africa and the unique LD characteristics could affect the performance of these algorithms. An unbiased comparison of these tools — other than the tests done by the tool authors —- is therefore imperative to determine their relative strengths and limitations.

2.3.2.2 Graph-based genotyping

To date, one of the bioinformatics advances that has not been extensively explored regarding the genotyping of pharmacogenes using HTS data is the implementation of genome graph-based variant detection algorithms. A genome graph structure (also termed as a pangenome) comprises nodes and edges which represent a linear reference genome that has been augmented with sample or population-specific variation (Ambler et al., 2019; Garrison et al., 2018; H. Li et al., 2020).

The Genome Reference Consortium (`https://www.ncbi.nlm.nih.gov/grc`) currently provides consensus linear haplotype assemblies for alignment of sequencing reads and calling variants. However, the popularly used linear human genome reference builds (GRCh37 and GRCh38) are not representative of all world populations and are missing up to 10% of genomic DNA of relevance to African populations in particular (Sherman et al., 2019), thus causing limitations in mapping accuracy especially for short-read NGS data. Some of the main read mapping issues include reference allele biases and read misalignment around indels (Garrison et

al., 2018; Martiniano et al., 2020). Futhermore, a number of patches and alternate contigs have been added especially to GRCh38 as more missing sequences continue to be discovered, thus making the data analysis involving alt-aware versus non-alt-aware alignment non-trivial (GATK team, 2021).

There is a growing interest in leveraging genome graph approaches for more accurate genomic variant analyses (Garrison et al., 2018; Rakocevic et al., 2019). This is because the genome graph structure enables numerous possible haplotypes from virtually all individuals in the population to be represented as paths in the reference (H. Li et al., 2020; Rakocevic et al., 2019). Therefore, genome graphs promote accurate read alignment due to the large number of haplotypes that can be represented in the graph, and faster genotyping for variants present in the graph (Eggertsson et al., 2017; Rakocevic et al., 2019). It is also easy to add novel variants/haplotypes to the existing pangenome. Initially, there were some concerns regarding the computational problem of handling such graph structures with dense haplotype information. However, enhancements in the form of graph indexes have facilitated the circumvention of this issue (Sirén et al., 2020). The utility of genome graph algorithms has already been demonstrated for complex loci such as HLA (Eggertsson et al., 2017). Given the high polymorphism associated with CYP genes, they present the perfect use case for pangenome-based genotyping approaches. It is unclear as to whether application of pangenome-based pharmacogene genotyping would be superior to existing allele calling approaches when applied to African genomes. Nonetheless, using tools that employ various genotyping approaches could be a better approach in order to capture various forms of novel and/or rare variation expected in previously unsampled African populations especially for *CYP2D6*, *CYP2B6* and *CYP2A6*.

2.3.3 Mining datasets comprising high-depth African genomes

A number of the previous pharmacogenetics studies that have elucidated the pharmacogene variation in continental African populations have been limited to relatively small sample sizes either by design or due to use of low-throughput experimental techniques, albeit with high accuracy (Masimirembwa et al., 1996; Matimba et al., 2009; Wright et al., 2010; Drögemöller, Plummer, et al., 2013; Radloff et al., 2013). There is increasing availability of high depth WGS data with more representation from understudied ethnolinguistic groups in Africa in part due to the aforementioned advancement in next generation sequencing technologies (Choudhury et al., 2020). The availability of high coverage African WGS data and the aforementioned allele calling bioinformatics tools provide the opportunity to study variation in important pharmacogenes such as *CYP2D6*, *CYP2B6* and *CYP2A6*, which have been found to be part of the 'inaccessible genome' when using low coverage HTS data (Drögemöller, Wright, et al., 2013). This will in turn lead to identification of common, rare and/or novel alleles and make the case for inclusion of

more African-specific variants on the more affordable array, panel-based and CNV genotyping platforms (Table 2.3). Population-scale characterisation of African-specific haplotypes would also provide a basis for future functional analysis assays and implementation of personalised medicine strategies. Some of the currently available high coverage datasets where continental African populations are represented include the following.

2.3.3.1 Data generated by the 1000 Genomes Project Consortium

The 1000 Genomes Project (Auton et al., 2015) has been one of the significant drivers in understanding genetic variation on a global scale over the past decade. By the time of its completion, the project had generated 2504 genomes representative of 26 populations, notably with 504 genomes of African individuals. This data was generated mostly using low-coverage WGS as well as deep exome sequencing and microarray genotyping. Although numerous studies have performed various analyses on these data, there have always been question marks regarding the variants observed in the *CYP2D6* region due to the low and/or non-uniform coverage (Nofziger et al., 2020). Recently, high-coverage whole genome sequence data has been generated for all the 2504 individuals (Byrska-Bishop et al., 2021). This data has been made publicly available, thus presenting an excellent resource for high-confidence interrogation of variations in the complex *CYP2D6*, *CYP2B6* and *CYP2A6* regions across multiple populations.

2.3.3.2 Data generated by the SAHGP and the H3A Consortium

In a pilot study, the Southern African Human Genome Programme (SAHGP) performed deep whole genome sequencing (33-51x) for 26 individuals to investigate the genomic diversity among southern African populations (Choudhury et al., 2017). The sample population included eight Coloured individuals and 16 black south-eastern Bantu-speakers. More recently, the Human Heredity and Health in Africa (H3A) Consortium (`https://h3africa.org`) has generated over 300 high coverage (~30x) WGS samples mainly from southern and west African populations. The bulk of this data has been archived in the European Genome-Phenome Archive (`https://ega-archive.org`). Initial genomic analyses using these data have identified a plethora of novel variations and multiple regions under selection pressures unique to Africa (Choudhury et al., 2020). It will therefore be interesting to infer potential novel pharmacogene star alleles and assess the relative distribution of known clinically relevant alleles across diverse African populations.

2.3.3.3 Polaris HiSeqX PGx Cohort

The HiSeqX PGx Cohort (Pratt et al., 2016) comprises 70 samples that have been characterised for variants of pharmacogenetic importance with different testing platforms at various laboratories as part of the CDC-based Genetic Testing Reference Materials Coordination Program (GeT-RM). These samples were collected from individuals (including 19 Yorubans) that were originally part of the HapMap (`https://www.genome.gov/10001688/international-hapmap-project`) and 1000 Genomes Projects (Auton et al., 2015). Illumina has recently generated high coverage whole genome sequence data for all 70 individuals in this cohort (`https://github.com/Illumina/Polaris`). Given that high-depth WGS data and gold standard star allele calls (Pratt et al., 2016) are available for these individuals, this dataset provides an invaluable resource for benchmarking pharmacogene genotyping platforms as well as star allele calling bioinformatics algorithms.

Summary

Precision medicine approaches are necessary in Africa and around the world to enhance the safety and efficacy of treatments for both communicable and non-communicable diseases — the burden of which is increasing in Africa. However, it is evident that information about pharmacogenomically relevant alleles is scarce across Africa. Recent availability of next generation technologies and high-depth genomic datasets from understudied African populations provide the opportunity to bridge this gap in knowledge. This thesis is focused on three pharmacogenes, i.e. *CYP2D6*, *CYP2B6* and *CYP2A6*, given their importance in drug metabolism. These three pharmacogenes are highly polymorphic and challenging to genotype especially when using computational approaches as reviewed in this chapter. The next chapter (Chapter 3) therefore presents a much-needed benchmarking analysis of the bioinformatics tools that have recently been developed to facilitate pharmacogene star allele calling from HTS data.

Chapter 3

A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: A focus on *CYP2D6* genotyping

The work described in this chapter was published as an open access article (see Appendix A and Appendix E) in *npj Genomic Medicine* under a Creative Commons Attribution 4.0 International License.

Twesigomwe D., Wright G.E.B., Drögemöller B.I., da Rocha J., Lombard Z. and Hazelhurst S. (2020). A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: A focus on *CYP2D6* genotyping. *npj Genomic Medicine* 5, 30. https://doi.org/10.1038/s41525-020-0135-2.

3.1 Summary

Genetic variation in genes encoding cytochrome *P450* enzymes has important clinical implications for drug metabolism. Bioinformatics algorithms for genotyping these highly polymorphic genes using high throughput sequence data and automating phenotype prediction have recently been developed. The *CYP2D6* gene is often used as a model during the validation of these algorithms due to its clinical importance, high polymorphism, and structural variations. However, the validation process is often limited to common star alleles due to scarcity of reference datasets. Additionally, there has been no comprehensive benchmark of these algorithms to date. We performed a systematic comparison of three star allele calling algorithms using 4618 simulations as well as 75 whole genome sequence samples from the GeT-RM project. Overall, we found that Aldy and Astrolabe are better suited to call both common and rare diplotypes compared to Stargazer, which is affected by population structure. Aldy was the best performing algorithm in calling *CYP2D6* structural variants followed by Stargazer, while Astrolabe had limitations especially in calling hybrid rearrangements. We found that ensemble genotyping, characterised by taking a consensus of genotypes called by all three algorithms, has higher haplotype concordance but it is prone to ambiguities whenever complete discrepancies between the tools arise. Further, we evaluated the effects of sequencing coverage and indel misalignment on genotyping accuracy. Our account of the strengths and limitations of these algorithms is extremely important to clinicians and researchers in the pharmacogenomics and precision medicine communities looking to haplotype *CYP2D6* and other pharmacogenes using high throughput sequencing data.

Keywords: *CYP2D6*, Genotyping algorithms, Phenotype prediction, NGS

3.2 Introduction

Genetic variation is known to influence the way in which individuals respond to therapeutics. Groups of variants that are inherited together, known as haplotypes, provide a basis for phenotype prediction and treatment decisions during pharmacogenomic testing. Accurately detecting functional haplotypes, popularly known as star (*) alleles, in clinically actionable pharmacogenes (e.g. cytochrome *P450* genes) is therefore crucial to the implementation of personalised medicine. In addition, determining pharmacogene star allele frequencies and consequently predicting the phenotypic landscape is a crucial part of large scale population genetic studies, which can help inform drug policy (Gaedigk et al., 2017).

Of the clinically-relevant pharmacogenes, *CYP2D6* is one of the most widely studied owing to its contribution to the Phase 1 metabolism of approximately 25% of clinically prescribed drugs, and high genetic variability within and between populations (Zanger and Schwab, 2013). Drugs metabolised by *CYP2D6* include various antidepressants, antipsychotics, anticancer agents (e.g. tamoxifen), and opioids (`https://drug-interactions.medicine.iu.edu/MainTable.aspx`, last accessed: 19-02-2020).

The highly polymorphic *CYP2D6* gene has nine exons and is located on chromosome 22q13.2 neighbouring two paralogous pseudogenes, *CYP2D7* and *CYP2D8*, which further complicates genotyping (Drögemöller, Wright, et al., 2013; Kimura et al., 1989). To date, the Pharmacogene Variation (PharmVar) Consortium (Gaedigk, Ingelman-Sundberg, et al., 2018) has catalogued over 130 different *CYP2D6* star alleles with varying levels of evidence (`https://www.pharmvar.org`, last accessed: 19-02-2020). The majority of *CYP2D6* star alleles are defined by specific combinations of single nucleotide polymorphisms (SNPs) and/or small insertions and deletions (indels) (collectively referred to as single/small nucleotide variants (SNVs) for the rest of this manuscript), which may either alter the function of the protein or be neutral. In addition, the *CYP2D6* gene locus contains a number of complex structural variants including full gene deletions, gene duplications and multiplications, as well as hybrid tandem rearrangements with the highly similar *CYP2D7* (Black et al., 2012; Gaedigk et al., 2010; Kramer et al., 2009). These allelic variants are graphically shown in Figure 3.1.

The complexities in the *CYP2D* gene locus present a significant challenge in accurately genotyping *CYP2D6*. Currently, gold standard genotyping involves a battery of methods e.g. quantitative PCR analysis for CNV detection, allele specific XL-PCR, SMRT sequencing, and Sanger sequencing. However, these techniques are expensive, laborious and unscalable, especially for large cohort studies (Gaedigk, 2013; Nofziger and Paulmichl, 2018; Yang et al., 2017). Further, although more economical, selectively targeting a subset of alleles through individually genotyping defining variants can have important clinical implications as inaccurate phenotype prediction is

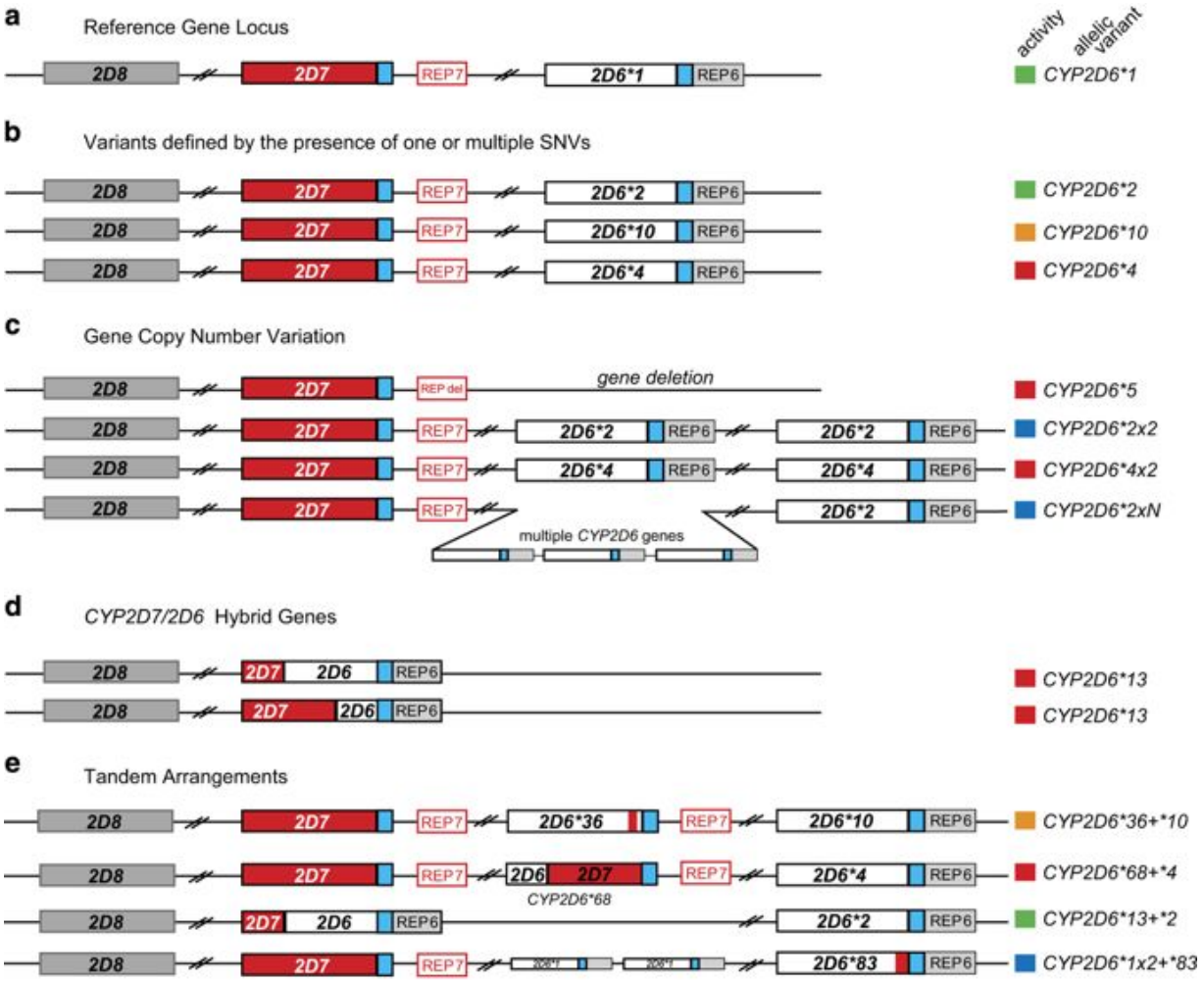


Figure 3.1: Graphical overview of the highly polymorphic *CYP2D6/2D7/2D8* locus, by Twist et al. (2016), licensed under Creative Commons CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`). No changes have been made to the figure content. (**a**) The relative position of the reference *CYP2D6*1* haplotype (white) to two non-functional paralogs, *CYP2D7* (red) and *CYP2D8* (grey) on the minus strand of Chromosome 22. REP6 and REP7 are paralogous, Alu-containing, 600-bp repetitive sequences found downstream of *CYP2D6* and *CYP2D7*, respectively. The blue boxes indicate identical unique sequences downstream of *CYP2D6* and *CYP2D7*. Notice the "spacer" (1.6-kb) separating REP7 from *CYP2D7* but none between *CYP2D6* and REP6. (**b**) Common *CYP2D6* star alleles defined by core single nucleotide variants (SNVs) (**c**) Examples of *CYP2D6* copy number variations and their functional annotation. (**d**) Examples of *CYP2D7/2D6* hybrid genes. (**e**) Common tandem rearrangements in the *CYP2D* gene locus. The activity level boxes on the right are coded; red for a non-functional haplotype, orange for decreased activity, green for fully functional reference activity, and blue for increased activity.

more likely to occur (Drögemöller, Wright, et al., 2013; Yang et al., 2017). The widespread availability of next-generation sequencing (NGS) characterised by high throughput parallel DNA sequence data generation provides a solution to the cost-limitations of experimental methods for genotyping *CYP2D6* and other pharmacogenes. The ever-decreasing cost of NGS facilitates whole genome sequencing (WGS) in addition to providing options for the development of gene

panels such as PGRNseq (Gordon et al., 2016).

In addition, NGS methods have the advantage of being unbiased, thus allowing for the detection of rare and novel variations, which are more likely to be deleterious (Wright et al., 2018) Although there are important benefits associated with NGS-based genotyping, these approaches present computational challenges, mainly relating to the alignment of short-read data in regions of high sequence similarity. This makes variant calling and haplotyping genes such as *CYP2D6*, that are located in regions of high sequence similarity, challenging.

To address these challenges, bioinformatics tools that have recently been developed for calling star alleles in *CYP2D6* and other highly polymorphic pharmacogenes using genome sequencing data and/or targeted-capture panels such as PGRNseq, include Astrolabe (formerly Constellation) (Twist et al., 2016), Aldy (Numanagić et al., 2018), Stargazer (S.-B. Lee, Wheeler, Patterson, et al., 2019), VCF Translator (Qiao et al., 2017), Cypiripi (Numanagić et al., 2015), and PharmCAT (Sangkuhl et al., 2020). These tools automate the detection of diplotype combinations based on PharmVar and the Pharmacogenomics Knowledgebase (PharmGKB) star allele catalogues thus facilitating clinical interpretation. The importance of these algorithms in *CYP2D6* phenotype prediction in clinical settings, and allele discovery in research studies, therefore cannot be overstated. However, to date, there is no comprehensive comparison of these tools. Astrolabe, Aldy, Stargazer, and PharmCAT are regularly maintained. However, PharmCAT does not perform star allele calling for *CYP2D6* directly, but rather uses Astrolabe's allele calling output for its unique clinical annotation step, which is based on current clinical implementation guidelines (Sangkuhl et al., 2020; Caudle et al., 2020).

We therefore provide a much-needed in-depth comparison of the performance of Astrolabe, Aldy, and Stargazer on a wide range of *CYP2D6* allelic variation using simulated data (entire *CYP2D* locus) and real data. The research participant-derived data used in the study are part of the Polaris pharmacogenomics cohort (`https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort#Pratt2016`) whose samples were originally collected as part of the HapMap and 1000 genomes projects (Auton et al., 2015). The Centers for Disease Control and Prevention (CDC)-based Genetic Testing Reference Material Coordination Program (GeT-RM) has characterised these samples through extensive orthogonal testing by various laboratories (Pratt et al., 2016; Gaedigk, Turner, et al., 2019). In doing so, GeT-RM has contributed invaluable reference materials for benchmarking *CYP2D6* genotyping approaches.

In addition, we evaluate the impact of read depth on recall of *CYP2D6* CNVs for each tool. Given the differences in the allele calling approach of these algorithms, we further analyse the efficacy of an ensemble genotyping approach involving all three tools using the real data.

Through this comparative analysis, we provide important information on the strengths and limitations of current pharmacogene star allele calling algorithms. Our recommendations will

serve as a valuable resource for the pharmacogenomics community regarding the use of these algorithms in clinical and research settings, and provide important data for developers aiming to improve the software or develop new ones. A summary of the main features of Astrolabe, Aldy, and Stargazer is given in Table 3.1.

Table 3.1: Properties of Astrolabe, Aldy, and Stargazer.

Tool	OS	Language	Central Feature	NGS Data	Input/Ref	Output
Astrolabe	Linux Mac	Java	Probabilistic scoring system	WGS [PGRNseq]	VCF BAM b37, b38	Diplotypes Subalelles Phenotype All novel SNVs
Aldy	Linux Mac	Python	Combinatorial framework	WGS PGRNseq	BAM b37	Diplotypes Subalelles Putative novel core variants
Stargazer	Linux Mac	Python	Statistical phasing	WGS PGRNseq	VCF GDF b37	Diplotypes Phenotype All novel SNVs

GDF, GATK-DepthOfCoverage Format; VCF, Variant Call Format; b37, Human Genome Reference Build 37; b38, Human Genome Reference Build 38; [], Square brackets indicate that support for PGRNseq data was not yet available for Astrolabe's CNV calling at the time of this study.

3.3 Methods

3.3.1 Ethics

Ethical approval was not required for this study as we used simulated and publicly available WGS data only.

3.3.2 Brief description for each algorithm

Detailed descriptions of all three algorithms are given in their original publications (Twist et al., 2016; Numanagić et al., 2018; S.-B. Lee, Wheeler, Patterson, et al., 2019).

Briefly, the Astrolabe (v0.8.7.0) algorithm uses probability scoring to identify the most likely *CYP2D6* diplotype match based on variant positions and zygosity from input variant call format (Danecek et al., 2011) (VCF) or gVCF files. Astrolabe also corrects for variant calling errors by using default or user-derived sensitivity and specificity of the variant calling approach. In addition, for WGS data, Astrolabe uses BAM (H. Li and Durbin, 2009) files for CNV calling

by comparing the alignment depth of coverage ratios of the *CYP2D6* exonic regions to the coverage across various control regions along chromosome 22. Astrolabe's output includes *CYP2D6* diplotypes for each sample, special notes on structural variants detected, suballeles, and diplotype activity information.

Aldy (v2.2.3) differs from Astrolabe by using a combinatorial approach to reconstruct the sequence content in each copy of a target polymorphic gene based on data from a sample's SAM (H. Li et al., 2009), BAM or CRAM (M. et al., 2011) file. Aldy leverages integer programming solvers such as Gurobi (`http://www.gurobi.com`), SCIP (`https://scip.zib.de`) and CBC (`https://developers.google.com/optimization`) for timely solution of problem instances (copy number estimation, major and minor star allele prediction) that would otherwise take an extended period of time to solve. In addition, Aldy performs a coverage normalisation step, which corrects for errors when calling alleles and CNVs from non-uniform coverage sequence data generated from target panels such as PGRNseq.

The Stargazer (v1.0.7) pipeline retrieves variant information from VCF files generated using GATK-HaplotypeCaller (Van der Auwera et al., 2013), then performs statistical haplotype phasing using Beagle (Browning and Browning, 2007), and thereafter assigns *CYP2D6* diplotypes based on haplotype matches with a reference star allele table. The Stargazer pipeline also includes supplementary algorithms for phasing variants that are not present in the reference VCF (default in v1.0.7 is the 1000 genomes reference panel). Of note, Stargazer also accepts phased VCF files, chip-generated VCFs, and provides the option for imputation. In addition, for structural variant detection, Stargazer requires read depth information for the *CYP2D6* and *CYP2D7* regions as well as a control gene region (e.g. *EGFR*, *RYR1*, and *VDR*), generated using GATK-DepthOfCoverage (Van der Auwera et al., 2013). The main output of Stargazer includes *CYP2D6* genotype data for each sample, interactive visual plots for CNVs, and predicted phenotype assignment based on the activity score system (Gaedigk et al., 2008; Gaedigk, Dinh, et al., 2018). In case a user has an automation platform installed, such as the Sun Grid Engine (SGE, `http://gridscheduler.sourceforge.net/htmlman/manuals.html`), Stargazer can be run using the BAM file as input.

3.3.3 Simulation of benchmark datasets

In order to compare the performance of Aldy, Astrolabe, and Stargazer in calling a large variety of *CYP2D6* haplotypes, we created simulated datasets with known *CYP2D6* diplotypes for unrelated individuals, following the approach described by Numanagić et al. (2015) with some modifications.

Maternal and paternal *CYP2D* locus sequences with respect to the human reference genome, GRCh37 (Chr22:42,518,000-42,555,000) were constructed separately for the simulation process.

For cases with no allele-defining structural variants, GATK FastaAlternateReferenceMaker (Van der Auwera et al., 2013) was used to mutate the *CYP2D6* portion of the maternal and paternal *CYP2D* sequences by giving *CYP2D6* haplotype VCF files from PharmVar (version 4.0.4) as input. Haplotype sequences with *CYP2D6/2D7* intron and/or exon conversions were constructed in a similar way since PharmVar gives details of all the SNVs arising from these structural changes in the VCF files.

CYP2D6 gene deletions were generated by creating breakpoints within the REP6 and REP7 regions as described by Gaedigk et al. (1991) and Steen et al. (1995) previously. *CYP2D6* duplications were generated by concatenating *CYP2D6* sequences containing desired mutations. For haplotypes with more complex structural variations such as *CYP2D6*61*, *CYP2D6*63*, and subvariants of *CYP2D6*13*, we replaced reference *CYP2D6* and/or *CYP2D7* portions of the *CYP2D* locus with haplotype sequences deposited in the NCBI Nucleotide database by Kramer et al. (2009), Black et al. (2012), and Gaedigk et al. (2010).

We simulated 101 bp paired-end reads (Illumina Hiseq2000) corresponding to maternal and paternal *CYP2D* haplotype sequences of each sample using the simNGS software (`https://www.ebi.ac.uk/goldman-srv/simNGS/`). simNGS replicates NGS-like noise intensity, base quality, and read error distributions by using runfiles trained from real sequencing experiments.

In order to account for the structural variant detection approaches used by Astrolabe and Stargazer, we simulated reads corresponding to CNV-neutral regions used by the algorithms and added them to the FASTQ files for each sample. The GRCh37 sequences (Chr22:19882000-19908000 and Chr22:44218000-44233000) contain control regions used by Astrolabe. For Stargazer, we used the *RYR1* gene locus (Chr19:38924340-39078204) as the control region. The CNV-neutral region that Aldy uses (Chr22:42547463-42548249) is covered by our simulation of the whole *CYP2D* locus.

Furthermore, to assess how various sequencing coverage affected Aldy, Astrolabe, and Stargazer's *CYP2D6* structural variant calling, we simulated read data with different coverage profiles that are frequently implemented in research and clinical settings ($\sim$30$\times$, $\sim$60x and $\sim$100x).

In total we generated 4618 simulations for Set 1, comprising only SNV-defined alleles. For Set 2, where each sample had at least one haplotype with allele-defining *CYP2D6* strucural variant(s), we had 3 replicates of 121 samples at different coverages as mentioned above. We excluded some *CYP2D6* star alleles from our analysis, including **32*, **99*, **110-*139*, either because they were not defined by all the three tools or they were still under curation by PharmVar. However, undefined suballeles were included to check for their effect on the tools' performance.

3.3.4 Real data

Regarding the real data, we used 75 samples from the CDC's GeT-RM project for which WGS data is publicly available. These samples were originally collected as part of the 1000 genomes and HapMap projects and were sequenced on Illumina HiseqX platform (2 x 150bp reads) (`https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort#Pratt2016`).

3.3.5 Variant calling

We aligned all our simulated read sets against the human reference genome (build 37) using BWA-MEM (H. Li and Durbin, 2009). We used the b37 reference in our analysis because it is supported by all three algorithms (Table 3.1). The GeT-RM WGS data were already in BAM format (b37) on retrieval. We obtained gVCF files from BAM files using GATK HaplotypeCaller followed by generation of VCF files with GATK GenomicsDB and GenotypeGVCFs (Van der Auwera et al., 2013).

3.3.6 *CYP2D6* star allele calling

We performed *CYP2D6* star allele calling separately for Aldy, Astrolabe and Stargazer using default parameters, as well as parameters recommended by the authors in cases of debugging and CNV calling. BAM files were used as input for Aldy, while Astrolabe required both BAM and VCF files. For Stargazer, we first generated depth of coverage (GDF) files, required for calling structural variants, using GATK3 DepthofCoverage. The GDF files were then used as input together with VCF files for each run. We included reference samples without structural variants to account for Stargazer's intersample normalisation.

For the GeT-RM samples, we determined consensus genotypes (ensemble) from the separate genotype calls of Astrolabe, Aldy and Stargazer using the 2 from 3 rule. Truth calls for these samples were obtained from previous publications by others (Gaedigk, Turner, et al., 2019; Pratt et al., 2016)

3.4 Results

3.4.1 Criteria

We evaluated Astrolabe, Aldy, and Stargazer using 3 criteria including;

1. **Haplotype concordance**: The percentage of total haplotype calls that are consistent with the ground truth. The haplotype true positive rate is also known as the analytic specificity.
2. **Diplotype (genotype) concordance**: The percentage of diplotype calls that are consistent with the ground truth (i.e. correctly calling both haplotypes for each sample).

3. **Phenotype concordance**: Percentage of assigned phenotypes (based on activity scores) concordant to the reference materials.

The results reported here are specific to Astrolabe (v0.8.7.0), Aldy (v2.2.3), Stargazer (v1.0.7) and the datasets (simulated and real) used in this study. The goal was to have datasets with a wide variety of catalogued *CYP2D6* alleles. The simulated data (4618 test cases) do not represent actual population frequencies. In comparison, the GeT-RM WGS data (75 samples) are representative of African, Admixed American, European, and Asian allele distributions.

3.4.2 Simulated Datasets

Set 1 of our simulations comprised all theoretically possible diplotype combinations (homozygous and heterozygous) derived from PharmVar-catalogued *CYP2D6* SNV-defined haplotypes. Set 2 comprised diplotypes with at least one structural variant (*CYP2D6* gene deletion (**5*), duplications, exon conversions and hybrid rearrangements i.e. *CYP2D6/2D7* and *CYP2D7/2D6* hybrids). These structural variations are comprehensively described by PharmVar (`https://www.pharmvar.org/gene-support/Variation_CYP2D6.pdf`). For a complete list of all the *CYP2D6* haplotypes used for generating simulations in this study, see Appendix C on page 216.

3.4.3 Performance based on Simulated Data

In Set 1 (SNV-defined alleles), the concordance of Astrolabe and Aldy for homozygous diplotypes was comparable as they correctly identified 142 (92%) and 148 (96%) diplotypes respectively out of 154 diplotypes (Table 3.2, Run 1). However, Stargazer had lower concordance (135 i.e. 88%) for the homozygous diplotypes. On further investigation, we found 18 cases where Stargazer prioritised reporting background functionally annotated alleles as the main result for haplotypes that have uncertain function. For example, Stargazer reported **58/*58* as **17/*17* given that **58* (unknown function) and **17* (decreased function) both have the T107I variant. In such cases, Stargazer reported the expected allele(s) ambiguously with other candidate solutions (see Table 3.2, and Table S3.1 on page 70 for additional details).

Calling some indel-defined alleles was challenging for all the tools due to left/right shifts in the alignments. These shifts emphasise challenges of using short-read NGS data for genotyping *CYP2D6* as they caused position discordance in the input data (BAM and/or VCF) and the algorithms' definitions (see Table 3.2 for examples of problematic indel-defined alleles). Discordant cases for Astrolabe and Aldy involved undefined suballeles as well (Table 3.2), resulting in ambiguous genotypes or miscalls. The concordance of Astrolabe and Aldy increased to 99% and 97% respectively after updating their allele databases with previously undefined suballeles (Table 3.2, Run 2).

Table 3.2: Concordance of Astrolabe, Aldy, and Stargazer on test cases (N=154) homozygous for SNV-defined star alleles at 30x. The complete list of genotypes called per sample is provided in Table S3.1 on page 70

	Astrolabe		**Aldy**		**Stargazer**
	Run 1	Run 2	Run 1	Run 2	
Match	142 (92%)	152 (99%)	148 (96%)	149 (97%)	135 (88%)
Mismatch	9 (6%)	1 (1%)	4 (3%)	3 (2%)	19 (12%)
Ungenotyped	3 (2%)	1 (1%)	2 (1%)	1 (1%)	0

Inconsistencies

Algorithm	**Challenges**	**Discordant cases**	**Notes**
Astrolabe	Undefined suballeles	**2.014*, **4.002*, **4.003*, **4.007*, **4.009*, **52.002*, **56.003*	Ambiguous calls or miscalls.
	Indel-defined alleles	*18.001	Position discordance with database.
Aldy	Undefined suballeles	*12.002, *71.003	Ambiguous calls or miscalls.
	Indel-defined alleles	*18.001, *20.001, *38.001	Position discordance with database.
Stargazer	Indel-defined alleles	*20.001	Position discordance with database.
	Alleles with unknown or uncertain function and/or rare alleles	18 alleles including **30.001*, **37.001*, **58.001*, **52.001*, **64.001*, **65.001*, and **73.001* (see Table S3.1)	Stargazer reports background alleles as the main haplotypes in these cases. Exact haplotypes (matching truthset) in the sample(s) are reported among other candidate haplotypes. (See discussion for details)

Run 1 and Run 2 indicate results before and after defining missing suballeles in the allele tables, respectively. There was only one run for Stargazer since it does not typically call suballeles.

We next considered tool concordance on all theoretically possible diplotype combinations of *CYP2D6* SNV-defined haplotypes with definitive or moderate PharmVar level of evidence (N=4560 i.e. 95 choose 2 plus 95). The main diplotypes called by all three tools are shown in Table S3.2 — see page 73. The concordance of Astrolabe and Aldy was again comparable (99% and 97%, respectively) while Stargazer was concordant for only (43%) of the possible diplotypes (Figure 3.2 on page 57). Stargazer's lower recall can be attributed to not only the aforementioned differences in reporting uncertain function alleles but also to the limitations of the statistical phasing process central to the Stargazer pipeline. For example, the **74/*48* combination involving a rare African-specific haplotype (**74*) and a rare East Asian-specific haplotypes (*48) was called as **1/*1* by Stargazer. Homozygous cases of these alleles are correctly called by Stargazer via manual phasing. However, rare heterozygous combinations are more difficult to call since population structure affects statistical phasing. Notably, some alleles are correctly called using Stargazer's "phasing by haplotype extension algorithm" that complements the statistical phasing (`https://stargazer.gs.washington.edu/stargazerweb/`). Astrolabe and Aldy do not do statistical phasing hence they are virtually not affected by population structure.

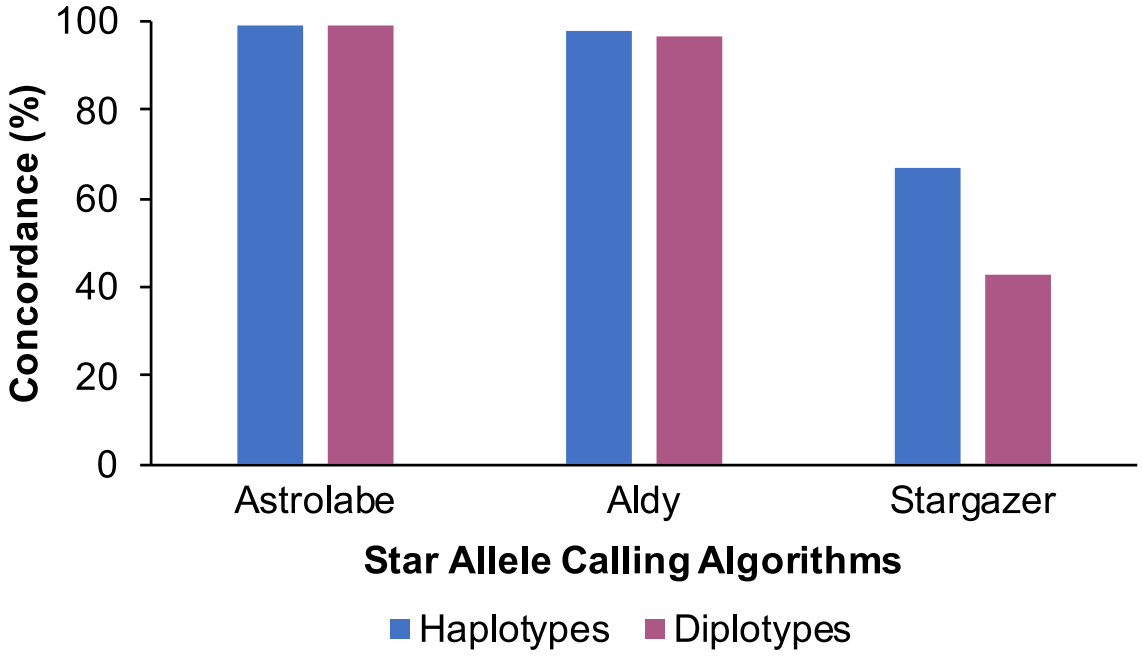


Figure 3.2: Concordance of Astrolabe, Aldy, and Stargazer for all theoretically possible *CYP2D6* diplotype combinations (homozygous and heterozygous) comprising SNV-defined haplotypes with PharmVar definitive or moderate level of evidence from our starting set. We left the allele databases of the 3 tools as is in order to examine the effect of the undefined suballeles on diplotype calling. The sequencing coverage for each test case was 30x and default parameters were used for each tool.

For all three tools, the concordance for haplotypes was much higher than that for diplotypes. This underscores the challenge of consistently calling both haplotypes for each sample in a highly polymorphic gene such as *CYP2D6*. We show the true extent of this later on with the real data that follows true population distributions.

All three tools accurately genotyped simulated samples with at least one *CYP2D6* full gene deletion allele (*5) (i.e. CN = 0, 1) as indicated in Table 3.3. The notable discordances were samples that contained a duplicated allele on the second haplotype (e.g. **5/*29x2* was called as **29/*29*). Astrolabe in particular also had challenges calling the **5/*5* diplotype (CN = 0). Astrolabe reported failed BAM quality control for the **5/*5* test cases. Running the algorithm by skipping quality control unsurprisingly yielded haplotype calls matching GRCh37 (i.e. **2M/*2M*, indicating absence of non-reference SNVs). Varying the coverage from 30x to 100x had very little effect on the *CYP2D6*5* recall of Aldy and Stargazer. However, Astrolabe detected 5 from 20 *CYP2D6*5* alleles at 60x compared to 10 from 20 at 30x and 100x. The miscalls were observed mainly in samples with CN = 1.

For copy number gains (CN>2), Aldy and Stargazer outperformed Astrolabe by detecting up

Table 3.3: Summary of correctly called structural variations by Astrolabe, Aldy, and Stargazer in Set 2. The complete list of genotypes called per sample is provided in Table S3.3 – Table S3.5 on pages 74 – 80

.

Set 2	Truth	Astrolabe			Aldy			Stargazer		
		30x	60x	100x	30x	60x	100x	30x	60x	100x
Full gene deletion (*5)	20	10	5	10	15	14	15	15	15	15
[a] Copy number gain	52	29	25	25	52	50	52	47	45	45
Resolved duplicated/ multiplicated alleles	61	4[c]	4[c]	4[c]	53	49	51	54	51	48
[b] Hybrids	107	3	8	8	72	66	67	31	30	28
Non-hybrid tandem	4	0	0	0	4	4	4	0	0	0
Ungenotyped (defaults)		0	0	0	1	3	2	18	16	19

[a] Due to gene duplications/multiplications.
[b] Collectively representing exon conversions and gene hybrids involving *CYP2D6* and *CYP2D7*.
[c] Determined only for homozygous allele cases.

to 50% more duplication/multiplication events. By default Aldy and Stargazer resolve copy number gains further to report which of the star alleles was duplicated or multiplicated. However, for our simulated test cases with high copy number, some discordant tandem arrangements were reported by Aldy and Stargazer. For example, Aldy genotyped the test case *CYP2D6*1x4/*2x8* as **1x3+*63/*2x7+*79*, while Stargazer genotyped it as **1x6/*34x6*. Genotypes with such high *CYP2D6* copy number are rare but possible as shown previously (Johansson et al., 1993). It is worth mentioning that Aldy was the only algorithm that resolved the **90+*1* non-hybrid tandem arrangements.

In contrast to Stargazer and Aldy, Astrolabe does not actively resolve duplicated/multiplicated alleles but rather generically indicates that a gene duplication has been detected. Similarly by default, Astrolabe does not distinguish between a duplication and a multiplication but rather generically indicates the presence of a gene duplication if copy number gain is detected. The concordance for duplications/multiplications was more or less equally good at 30x, 60x and 100x for Aldy and Stargazer (Table 3.3). For Astrolabe, there was no evidence to suggest that increasing coverage from 30x to 60x or 100x increases recall of duplications. This could be due to limitations of using short-read NGS data for interrogating the complex *CYP2D6* region.

Regarding gene exon conversions and hybrid rearrangements, Aldy consistently called more of such alleles at 30x (72 out of 107) compared to Stargazer (31 out of 107) and Astrolabe (3 out of 107). Astrolabe consistently called **82* at 30x, 60x, and 100x unlike Aldy and Stargazer, and then consistently called the **68+*4* tandem only at 60x and 100x. As observed for duplication events, increase in coverage from 30x to 60x and 100x did not considerably affect Aldy and

Stargazer's performance for calling hybrids (See Table 3.3 above, and Table S3.3 – Table S3.5 on pages 74 – 80 for additional details).

3.4.4 Comparison of *CYP2D6* star allele calling performance on real data

To evaluate the genotyping accuracy of Aldy, Astrolabe, and Stargazer on real data, we applied all three algorithms to 75 ethnically diverse samples from the GeT-RM project. GeT-RM previously characterised 137 DNA samples from Coriell cell lines for 28 genes, including *CYP2D6*, through targeted genotyping by selected laboratories. (Pratt et al., 2016) GeT-RM has recently re-characterised these 137 samples and genotyped 42 new samples to ascertain complex rearrangements and rare variants in *CYP2D6* (Gaedigk, Turner, et al., 2019). Of the combined 179 samples, 75 currently have publicly available high coverage WGS data (see subsection 3.6 for details). We used published consensus *CYP2D6* genotypes (Gaedigk, Turner, et al., 2019; Pratt et al., 2016) for these samples as the Gold Standard calls.

When applied to the GeT-RM WGS samples, both Astrolabe and Stargazer assigned diplotypes for 75 of 75 samples. One sample could not be genotyped by Aldy with default parameters. Stargazer had the highest genotype concordance (88% – 66 of 75 diplotypes) to the ground truth followed by Aldy (87% – 65 of 75 diplotypes) and then Astrolabe (71% – 53 of 75 diplotypes). The haplotype concordance followed a similar trend i.e. 93% (139 of 150 haplotypes), 89% (134 of 150 haplotypes), and 82% (123 of 150 haplotypes) for Stargazer, Aldy, and Astrolabe respectively (Figure 3.3a).

Notably, the ensemble genotyping approach (2 from 3 rule) had the highest diplotype concordance (93% – 70 of 75 diplotypes) and haplotype concordance (95% – 143 of 150 haplotypes).

Samples with *CYP2D6* CNVs were more challenging to call for all the algorithms notably Astrolabe (15 of 75 samples) (see Figure 3.3b). Astrolabe particularly had challenges calling **36*, a *CYP2D6*/2D7 hybrid gene that frequently occurs singly or in tandem with **10*. Tandem rearrangements involving *36 present in the GeT-RM dataset include **36x2*, **36+*10*, and **36x2+*10*. Aldy and Stargazer consistently detected the more common **36+*10* arrangement but could not call all copies correctly for the more complex **36x2* and **36x2+*10* arrangements (See Table S3.6 on page 83 for detailed genotype calls).

All the three tools consistently called the *68+*4 tandem arrangement except for one case where Astrolabe called sample NA21781 as **2/*4* (possible duplication) instead of **2x2/*68+*4*. *CYP2D6*68* is a non-functional *CYP2D6*/2D7 hybrid gene with an intron 1 breakpoint. The **68+*4* haplotype is therefore non-functional as **4* is also a non-functional allele.

As shown in Figure 3.3c, haplotype calls by one tool that are not confirmed by either of the other two tools are mostly incorrect. The caller overlap between Aldy and Stargazer was

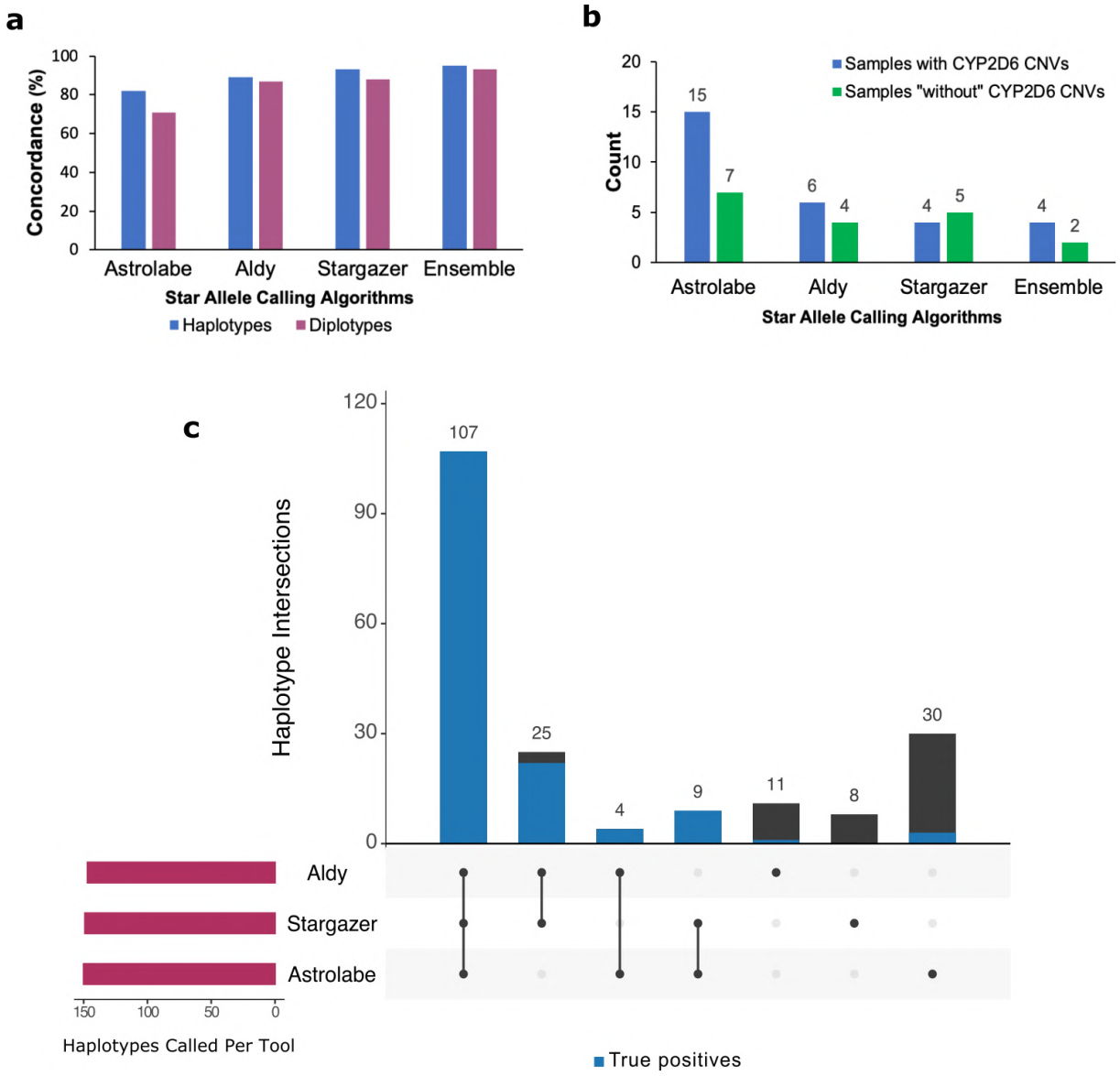


Figure 3.3: Performance of Astrolabe, Aldy, and Stargazer on 75 WGS GeT-RM samples. (**a**) Concordance of each algorithm to the GeT-RM consensus calls. (**b**) Cases with discordant genotypes. The green color represents samples "without" *CYP2D6* CNVs (most of the samples may have the intron 1 conversion) while the blue color represents samples with allele-defining *CYP2D6* CNVs. (**c**) Overlap of haplotypes called by each algorithm. As shown ensemble call sets have high concordance. Haplotypes called by one tool but not confirmed with any of the others have low true positive rate.

more prominent than either tool's intersection with Astrolabe because of Astrolabe's lower recall for structural variations. The ensemble genotyping approach reduced the number of discordant genotype calls especially for samples "without" *CYP2D6* CNVs (many of these have haplotypes with the neutral intron 1 conversion). However, some alleles defined by *CYP2D6* CNVs (especially aforementioned **36* hybrid arrangements) proved to be challenging even with ensemble calling (Figure 3.3b). Additionally, the ensemble genotyping approach had 2

ambiguous calls (from samples NA19908 and NA18540) arising from discrepant calls from all 3 tools (Table S3.6 on page 83).

3.4.5 Comparison of *CYP2D6* phenotype prediction concordance

We next evaluated the clinical accuracy of Astrolabe, Aldy, Stargazer, and the ensemble approach. We followed the current consensus clinical implementation guidelines (Caudle et al., 2020; Nofziger et al., 2020) to assign activity scores corresponding to the genotype calls and compared them to the GeT-RM consensus reference. All the three algorithms had much higher phenotype concordances compared to their diplotype concordances (Figure 3.4). This is due to the fact that even though some samples may not be accurately genotyped, the activity score or phenotype group of the reported diplotype could be the same as that of the truth call. For instance, sample NA18565 was called as **10/*36+*10* (Stargazer and Aldy) and **10/*10* (Astrolabe) instead of **10/*36x2* (truth call). All these genotypes denote an intermediate metaboliser phenotype. Aldy (95%) and Stargazer (95%) had comparably higher phenotype concordance than Astrolabe (89%) because they were able to consistently genotype more samples with *CYP2D6* CNVs as mentioned earlier. Notably, the ensemble approach had the highest (96%) phenotype concordance and it was comparable to its diplotype concordance (93%). In some cases, we obtained phenotype "no calls" due to ungenotyped samples e.g. for Aldy(1) and the ensemble approach (2 ambiguous cases).

3.5 Discussion

Accurately calling star alleles in *CYP2D6* is crucial to inferring phenotype from genotype. However, *CYP2D6* is one of the most challenging genes to genotype largely because of the complexity of its genomic locus. The limitations of using short-read NGS data compound the problem even further. Herein, we have benchmarked three algorithms that use largely different approaches to call *CYP2D6* star alleles from high throughput sequence data.

We used simulations to test the performance of each algorithm on a wide variety of *CYP2D6* allelic variation i.e. 154 SNV-defined haplotypes (many have the *CYP2D7* intron 1 conversion), and over 50 different structural variant arrangements (Appendix C on page 216). We also assessed the accuracy of each algorithm using real WGS data from CDC-based GeT-RM project.

Overall, Astrolabe and Aldy call more combinations of SNV-defined *CYP2D6* star alleles (homozygous and heterozygous) compared to Stargazer. Stargazer faces a challenge in calling rare (frequency $< 0.1\%$) SNV-defined alleles due to its reliance on statistical phasing. Although Stargazer has supplementary algorithms to phase rare variants that are not present in the reference panel, we found that they work best in homozygous cases but have lower accuracy in heterozygous cases. Stargazer also had discordances due to its difference in reporting alleles with unknown or

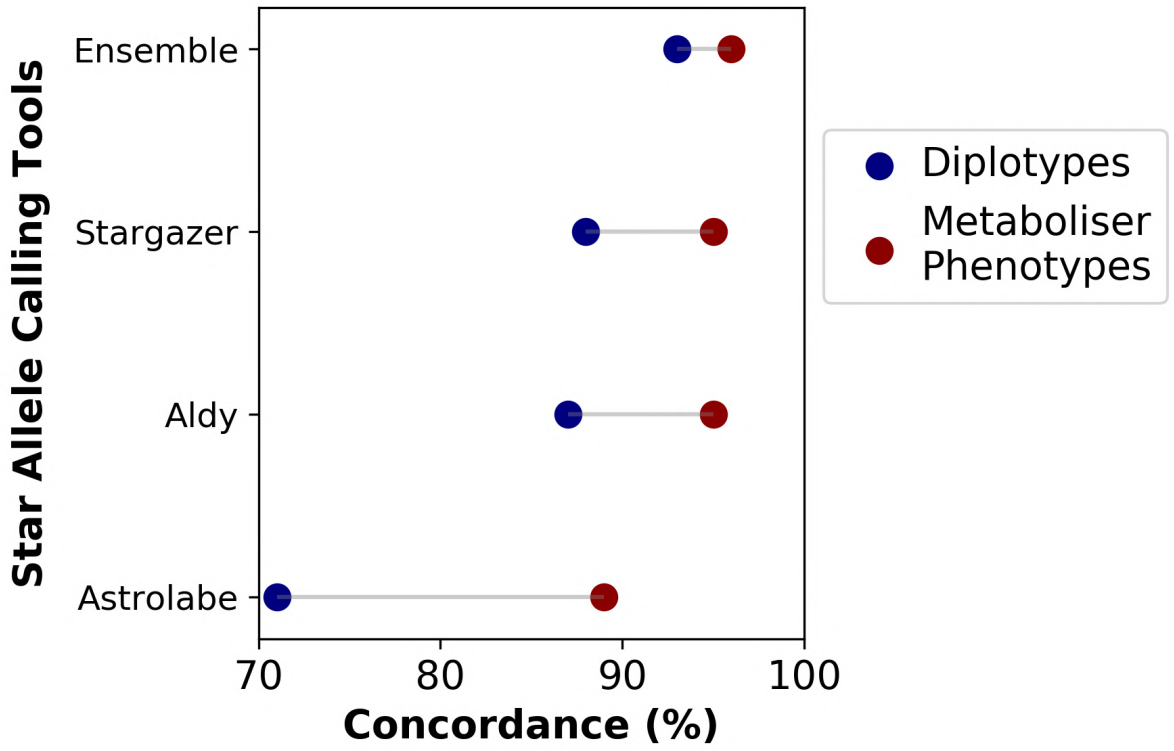


Figure 3.4: Comparison between the *CYP2D6* diplotype concordance and the phenotype prediction concordance for each algorithm based on 75 GeT-RM samples and the Activity Score system.

---

uncertain function. For such alleles, Stargazer currently prioritises to report background alleles for which the function is known as the main haplotypes. However, it is difficult to infer the exact alleles in the sample as they are reported in a separate column ambiguously with other candidate alleles. Notably, homozygous cases of these alleles are assigned unfounded phenotypes. For example, **58/*58* (unknown metaboliser status) is called as **17/*17* (intermediate metaboliser) by Stargazer while **58* is reported among candidate alleles for haplotype 1 and 2.

Notably all three algorithms depend on the comprehensive definition of alleles (and suballeles especially for Astrolabe and Aldy) to avoid ambiguous calls and miscalls. Indel-defined star alleles are a major challenge for all three algorithms especially where the core SNV occurred in a repeat region causing the aligner to pick the left-most possible position in the alignment (BAM) file. Core indels for **18*, **20*, **38*, **40* and **42* all appear in unexpected positions due to this problem. Miscalling these alleles could have potential clinical implications. Even though the three algorithms have corrected for some of these changes in their allele definitions, the complexity of their locus can affect coverage and variant quality scores. Of note, other alleles with indel core variants such as **3*, **6*, **9*, **19*, and **21* were consistently called.

Regarding the calling of *CYP2D6* structural variants, duplications/multiplications and deletions

were considerably easier to call compared to hybrid rearrangements. Aldy had the highest concordance followed by Stargazer, and then Astrolabe. Notably, Astrolabe does not distinguish between duplications and multiplications (i.e. identifies all scenarios as duplications), and also doesn't resolve the exact duplicated gene copy (or copies) for samples where it detects copy number gains. This could affect phenotype prediction depending on the diplotype combination in a sample. Astrolabe also had challenges detecting the **5/*5* diplotype (CN = 0) probably because its current model for structural variant calling was not trained on enough cases with zero copy number. In addition, for hybrid rearrangements, Astrolabe consistently detected only **82* (has an exon 2 conversion) and **68+*4* (at 60x and 100x). As with the duplication events, the effect of Astrolabe's inconsistency in calling hybrid rearrangements on phenotype prediction depends on the entire diplotype combination of an individual.

Stargazer had the intermediate performance for calling structural variants of the three tools. It was able to phase CNVs and detect more hybrid rearrangements compared to Astrolabe while also producing visual plots (Figure S3.1) of the structural variants in the process. However, Stargazer had the highest number of indeterminate (i.e. ungenotyped) diplotypes for samples with structural variants. Notably, Stargazer still produced copy number and allele fraction plots for these samples. However, ungenotyped calls can be a major hindrance to automated phenotype prediction in practice.

Aldy performed best in calling *CYP2D6* structural variants. However, as with Astrolabe and Stargazer, it had challenges in reconstructing complex hybrid rearrangements such as *CYP2D6*79+*68+*4* and **79+*68+*4x2*, and it also incorrectly indicated the presence of hybrids such as **63* and **79* for samples with high copy number. Although high copy number diplotypes such as **1x4/*2x8* are rare, it is important to examine whether they can be picked up by these algorithms. This is particularly important given that some populations are largely understudied and may have individuals with such alleles.

To evaluate the performance of Astolabe, Aldy, and Stargazer on real data, we used 75 sequences from the 1000 Genomes Project for Coriell reference samples that have been recharacterised through the GeT-RM project. Astrolabe and Stargazer were able to call diplotypes for 100% of the samples, while one sample could not be diplotyped using Aldy's default parameters. However, the concordance of Aldy and Stargazer was comparably higher than that for Astrolabe mainly due to their superiority in calling *CYP2D6* CNVs. Although the 75 GeT-RM WGS samples used for this analysis depicted the complexity and diversity in *CYP2D6* pharmacogenomic variation, they did not include a number of rare alleles. Future availability of WGS data for the remaining 104 GeT-RM reference materials will provide the opportunity to test these algorithms against a more comprehensive distribution of *CYP2D6* star alleles.

As with the simulated data, gene deletions as well as duplications/multiplications were more

consistently called compared to hybrid rearrangements. However, Astrolabe missed more deletions and duplications than Aldy and Stargazer. This could be attributed to Aldy and Stargazer's coverage normalisation strategies during structural variant detection that are not utilised by Astrolabe.

Regarding hybrids, the *CYP2D6*68+*4* non-functional tandem arrangement was consistently called by all three algorithms. Notably, Astrolabe could not differentiate *CYP2D6*36* from **10*. On the other hand, Aldy and Stargazer consistently called the **36+*10* arrangement but could not differentiate more complex arrangements such as **36x2*, **36x2+*10*. This can be attributed to incomprehensive definition of these haplotypes within the algorithms and also the limitations of using short-read NGS data.

We also assessed the overlap of Astrolabe, Aldy, and Stargazer calls as well as the practical application of ensemble *CYP2D6* star allele calling. Simple ensemble calling could be useful in obtaining high-confidence *CYP2D6* star allele calls both in the clinical setting and in large scale population studies. Through obtaining high-confidence star allele calls, the number of samples requiring experimental validation (e.g. those found to have putative novel alleles and structural variants) would be reduced to a smaller subset of the study population. We generated the ensemble call set by taking a consensus of Aldy, Astrolabe and Stargazer allele calls for each sample on a 2 from 3 rule. We obtained 100% concordance for haplotypes (107) that were called by all three algorithms, and only 6 discordant genotypes compared to Astrolabe (22), Aldy (10), and Stargazer (9) from 75 samples. Based on the challenges and limitations of each algorithm and of using short-read NGS data, we recommend the use of all three algorithms to perform highly accurate ensemble calling. The ensemble calling is also practical as the typical runtime for each of three algorithms from our workflows was less than 1 minute per sample while using 4 CPU cores.

Although it is quite straightforward to obtain high-confidence SNV-defined star allele calls using ensemble calling, it is more challenging to get unambiguous high-confidence diplotype calls for individuals with complex structural variations especially the hybrid rearrangements. This follows from the differences in accuracy of the three algorithms for different forms of *CYP2D6* allelic variation described above. A worry with ensemble calling in the clinic would be how to resolve the discrepancies among the tools. These samples would require orthogonal testing to resolve the number and exact sequence content of their *CYP2D6* copies. Further work is required to build an automated ensemble allele calling and phenotype prediction pipeline. This will heavily depend on achieving uniformity in the precision of reporting star alleles and suballeles by all the algorithms in their future updates.

All three algorithms had much higher phenotype concordance compared to their diplotype concordance implying that their clinical accuracy does not necessarily match their genotyping

accuracy. Notably, the ensemble genotyping approach had comparably high phenotype and diplotype concordance in contrast to Astrolabe, Aldy, and Stargazer. This implies that ensemble genotyping followed by phenotyping, minimises incorrect predictions of individual *CYP2D6* diplotypes as well as metaboliser status for which the three algorithms could be prone if used singly.

The pros and cons of Astrolabe, Aldy, Stargazer, and the ensemble approach based on our analysis are summarised in Table 3.4 on page 67.

Our approach had some limitations. Firstly, we used simulated data for most of the benchmarking especially for rare alleles and structural variants. Therefore, we did not capture the alterations in performance caused by some of the features present in real data for these cases. However, publicly available reference datasets with validated *CYP2D6* diplotype calls are limited and they also have drawbacks as they do not cover a wide variety of *CYP2D6* allelic variation. The GeT-RM reference materials are more or less comprehensive; however, public WGS data is not yet available for all the samples or biobank scale cohorts. The different relative performances between the algorithms on simulated data and real data are attributed to the different distribution of alleles (i.e. fewer rare alleles in the real data). Our exhaustive simulation enabled us to test the genotyping algorithms on a wider array of *CYP2D6* variations. On the other hand, the small cohort of real world patients was very helpful for evaluating results from the algorithms against those obtained by using orthogonal techniques. The 75 GeT-RM samples are representative of 24 *CYP2D6* major star alleles (17 SNV-defined haplotypes plus 7 structural variants) compared to the 204 haplotypes (154 SNV-defined haplotypes plus 50 structural variants) represented in our simulated data.

We also did not compare the performance of the three tools across simulated or real NGS data from targeted custom-capture panels such as PGRNseq. These panels greatly reduce the cost of sequencing and enhance genotyping turnaround time. Further, we did not compare the influence of the aligner, BWA-MEM versus other aligners such as GSNAP (Wu and Nacu, 2010) and novoalign (`http://www.novocraft.com`).

It is important to note that these algorithms face a major challenge of genotyping individuals with novel SNVs which may either be neutral or allele-defining (based on unpublished data). We did not test for this in this study as the reference materials used did not have novel star alleles. Ideally, for such cases, the two background star alleles would be called but the novel SNV(s) would be ambiguously assigned to either haplotype. Experimental validation using approaches such as Sanger sequencing, XL-PCR and/or long-read sequencing (e.g. PacBio, and Oxford Nanopore) would then be required to unequivocally determine the phase of these novel SNVs. Caution also needs to be taken in cases of complex and/or novel CNVs as they might be misreported by the algorithms. Nonetheless, combining short-read NGS (especially WGS) and

tools such as Astrolabe, Aldy, and Stargazer (or their ensemble) is crucial to characterisation of pharmacogene variation in a cost-effective manner by reducing the number of cases that require the more expensive experimental genotyping approaches.

Even though our comparison is focused on *CYP2D6*, we acknowledge that Astrolabe, Aldy, Stargazer, and PharmCAT (not included in the analysis) are extremely useful in genotyping other pharmacogenes for example *CYP2C9*, *CYP2C19*, *CYP2A6*, *CYP2B6*, *CYP3A4*, and *CYP3A5* as shown previously by others (S.-B. Lee, Wheeler, Thummel, et al., 2019; Numanagić et al., 2018; Sangkuhl et al., 2020; Twist et al., 2016) (See also Table S3.7).

Through this benchmarking study, we have examined the major differences as well as strengths and limitations of three algorithms that perform the challenging task of genotyping *CYP2D6* and other highly polymorphic pharmacogenes using short-read NGS data. Some of the inaccuracies reported were due to sequencing noise in the simulated and real data. This supports the notion that clinical-grade NGS data (preferably high coverage WGS) is required for effective use of these algorithms in the clinical setting. It is also clear that more *CYP2D6* pharmacogenetic studies especially in understudied populations would add to the completeness of the allele catalogue in PharmVar and in the allele definitions of these *in silico* genotyping algorithms hence enhancing their utility across populations.

3.6 Data Availability Statement

The real datasets used in this study are publicly available from the European Nucleotide Archive (`https://www.ebi.ac.uk/ena/data/view/PRJEB19931`). Samples NA12878, NA12891, and N12892 can also be downloaded from the 1000 genomes project repository (`https://www.internationalgenome.org`). Metadata for all samples can be found at `https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort#Pratt2016`.

The VCF files used to create hg19 consensus haplotype sequences for simulations are publicly available from the PharmVar database (`https://www.pharmvar.org/gene/CYP2D6`). All the fasta sequences used to generate *CYP2D6* CNV simulations are available at `https://github.com/twesigomwedavid/CYP2D6-gt-algorithms`.

3.7 Code Availability Statement

The code used for this analysis is publicly available at `https://github.com/twesigomwedavid/CYP2D6-gt-algorithms`.

Table 3.4: Pros and cons of current *CYP2D6* genotyping algorithms

Algorithm	Advantages	Disadvantages
Astrolabe	High accuracy for calling catalogued SNV-defined alleles. Supports hg19 and hg38. Performs variant calling error correction. Easy to run (one-liner command). Performs automated phenotype prediction.	Lower recall for *CYP2D6* CNVs and hybrid rearrangements compared to Aldy and Stargazer. Prone to ambiguous calls and miscalls if alleles/subalelles are not comprehensively defined. Does not discriminate duplicated alleles in a sample with copy number gain (generically reports presence of a duplication). Does not distinguish between duplication events and multiplication events.
Aldy	High accuracy for calling catalogued SNV-defined alleles. Highest accuracy of the three tools in calling *CYP2D6* structural variations. Easiest to run of the three tools as it requires only the BAM file on any system (even a normal laptop).	Aldy supports only hg19 as of v2.2.3. Prone to ambiguous calls and miscalls if alleles/subalelles are not comprehensively defined. Prone to some erroneous calls when using default arguments due to sequencing noise.
Stargazer	Supports various file inputs and imputation. Supports user-defined references for phasing and/or phased VCF input. Provides tools for viewing metrics of key variants in a sample. Outputs coverage plots for visually examining change-points for samples with CNVs. Performs automated phenotype prediction. Easy to run (one-liner command).	Low recall for rare alleles especially in heterozygous combinations. Affected by linkage disequilibrium as it has to do statistical reference-based phasing. Prone to "no calls" for complex hybrid arrangements/combinations. Stargazer v1.0.7 supports only hg19. Does not call subalelles as of v1.0.7. Has inconsistent reporting of alleles with uncertain function (reports background functionally annotated alleles for these cases as of v1.0.7).
Ensemble	Comparably high diplotype/haplotype concordance. Resolves single tool deficiencies.	Difficult to resolve complete discrepancies between the genotypes of the three tools. Prone to ambiguous calls especially for structural variations. No automated pipeline as of now as the three algorithms are being updated on a regular basis and the reporting of alleles/subalelles is non-uniform.

3.8 Acknowledgements

The authors thank the developers of Astrolabe, Aldy, and Stargazer for technical assistance with running the algorithms. The authors acknowledge the Genetic Testing Reference Material Coordination Program (GeT-RM) for their contribution of characterised *CYP2D6* reference materials and whole genome sequencing data for the corresponding samples. Special appreciation goes to Shaun Aron and Phelelani Mpangase for their support during the bioinformatics analysis at the Sydney Brenner Institute for Molecular Bioscience. The authors would also like to thank Annette Gross from the Clinical Pharmacology Modelling and Simulation Department in GlaxoSmithKline's Research and Development organisation for her perspectives on this work. This work was primarily funded through a grant by GlaxoSmithKline Research and Development Ltd to the Wits Health Consortium. GSK was not responsible for the study design and the views and opinions expressed in this paper are not necessarily those of GSK. D.T and J.dR are partially funded by the South African National Research Foundation.

3.9 Competing Interests

The authors declare no competing interests.

3.10 Author Contributions

Study concept and design: D.T., S.H., Z.L., J.dR. Acquisition, analysis or interpretation of the data: D.T., S.H., G.W., B.D. Drafting of the manuscript: D.T., G.W., B.D. Critical revision of the manuscript for important intellectual content: G.W., B.D., J.dR, S.H., Z.L. Obtained Funding, Administrative, Technical or Material Support: S.H. Study Supervision: S.H., Z.L.

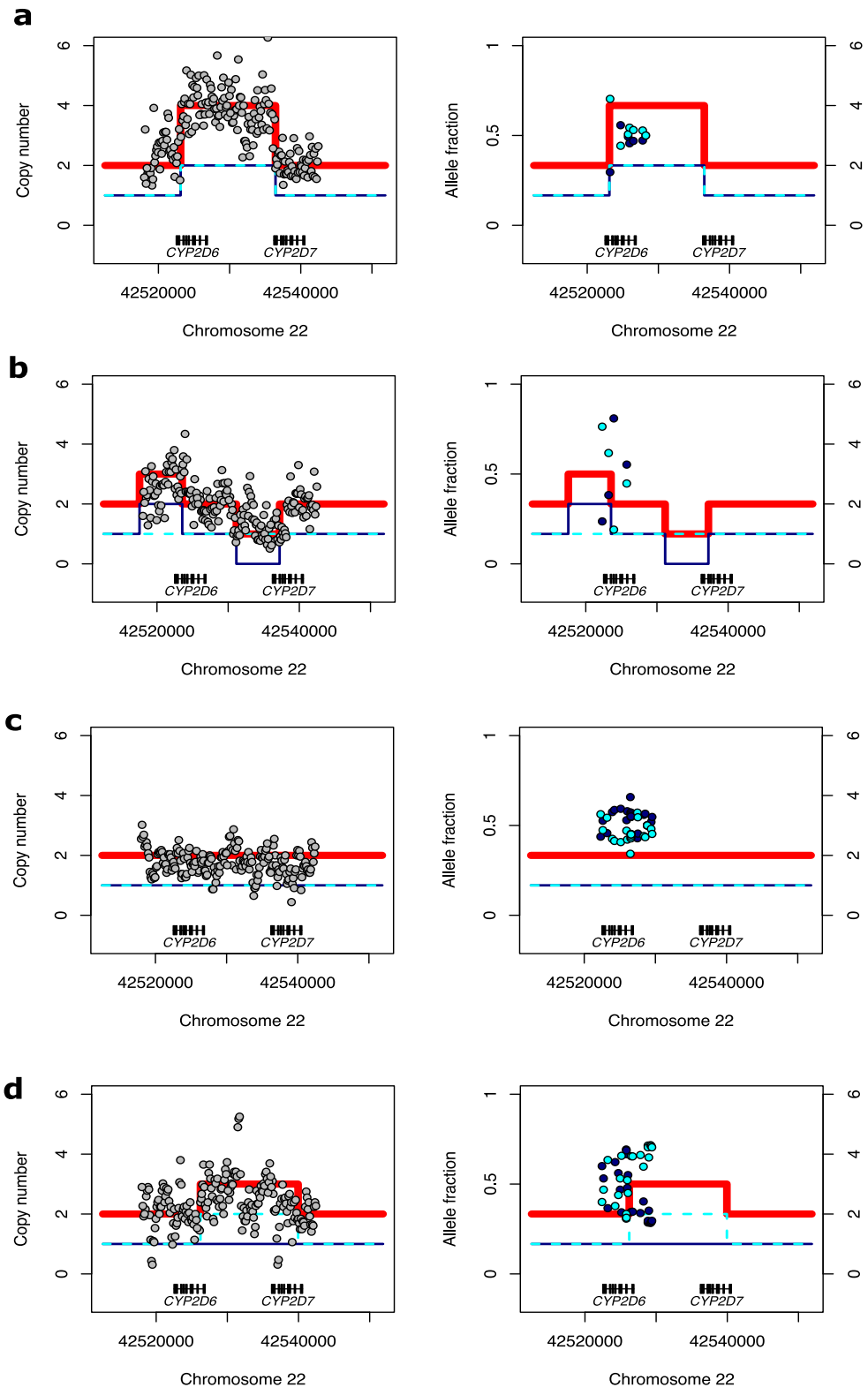


Figure S3.1: Examples of Stargazer plots for structural variants in Set 2. (a) Shows SIM119 with the **36+*10/*36+*10* diplotype. Notice the duplication and the exon 9 conversion deduced for both haplotypes (sky blue and navy blue dotted lines) (b) Shows SIM198 with a *CYP2D7/2D6* hybrid rearrangement (**13+*2*). The other haplotype is *CYP2D6*39*. (c) Shows NA12892 with the **2/*3* diplotype (included for comparison). (d) Shows NA12878 with the **3/*68+*4* diplotype. Notice the *CYP2D6* to *CYP2D7* gene conversion from intron 1 onward shown by the dip in depth of coverage.

Table S3.1: Stargazer, Astrolabe, and Aldy results for simulated homozygous star alleles (N=154)

Test Case	Truth diplotype	Stargazer				Astrolabe		Aldy	
		hap1_main	hap2_main	hap1_candidate	hap2_candidate	Run 1	Run 2	Run 1	Run 2
CYP2D6_1.001_CYP2D6_1.001	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.002_CYP2D6_1.002	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.003_CYP2D6_1.003	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.004_CYP2D6_1.004	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.005_CYP2D6_1.005	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.007_CYP2D6_1.007	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.008_CYP2D6_1.008	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.014_CYP2D6_1.014	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.019_CYP2D6_1.019	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_1.025_CYP2D6_1.025	*1/*1	*1	*1	*1	*1	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_2.001_CYP2D6_2.001	*2/*2	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	*2/*2	*2/*2	*2/*2	*2/*2
CYP2D6_2.002_CYP2D6_2.002	*2/*2	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	*2/*2	*2/*2	*2/*2	*2/*2
CYP2D6_2.003_CYP2D6_2.003	*2/*2	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	*2/*2	*2/*2	*2/*2	*2/*2
CYP2D6_2.004_CYP2D6_2.004	*2/*2	*2D	*2D	*2D,*34	*2D,*34	*2/*2	*2/*2	*2/*2	*2/*2
CYP2D6_2.010_CYP2D6_2.010	*2/*2	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	*2/*2	*2/*2	*2/*2	*2/*2
CYP2D6_2.011_CYP2D6_2.011	*2/*2	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	*2/*2	*2/*2	*2/*2	*2/*2
CYP2D6_2.013_CYP2D6_2.013	*2/*2	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	*2/*2	*2/*2	*2/*2	*2/*2
CYP2D6_2.014_CYP2D6_2.014	*2/*2	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	[mac]	*2/*2	*2/*2	*2/*2
CYP2D6_2.017_CYP2D6_2.017	*2/*2	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	*2/*2	*2/*2	*2/*2	*2/*2
CYP2D6_3.001_CYP2D6_3.001	*3/*3	*3	*3	*3	*3	*3/*3	*3/*3	*3/*3	*3/*3
CYP2D6_3.002_CYP2D6_3.002	*3/*3	*3	*3	*3	*3	*3/*3	*3/*3	*3/*3	*3/*3
CYP2D6_4.001_CYP2D6_4.001	*4/*4	*4	*4	*4,*10,*39	*4,*10,*39	*4/*4	*4/*4	*4/*4	*4/*4
CYP2D6_4.002_CYP2D6_4.002	*4/*4	*4	*4	*4,*10	*4,*10	*74/*74	*4/*4	*4/*4	*4/*4
CYP2D6_4.003_CYP2D6_4.003	*4/*4	*4	*4	*4,*10,*39	*4,*10,*39	*10/*10	*4/*4	*4/*4	*4/*4
CYP2D6_4.004_CYP2D6_4.004	*4/*4	*4	*4	*4,*10,*39	*4,*10,*39	*4/*4	*4/*4	*4/*4	*4/*4
CYP2D6_4.007_CYP2D6_4.007	*4/*4	*4	*4	*4,*10,*39	*4,*10,*39	*10/*10	*4/*4	*4/*4	*4/*4
CYP2D6_4.009_CYP2D6_4.009	*4/*4	*4	*4	*4,*10	*4,*10	*74/*74	*4/*4	*4/*4	*4/*4
CYP2D6_4.012_CYP2D6_4.012	*4/*4	*4	*4	*4	*4	*4/*4	*4/*4	*4/*4	*4/*4
CYP2D6_4.014_CYP2D6_4.014	*4/*4	*4	*4	*4,*10,*39	*4,*10,*39	*4/*4	*4/*4	*4/*4	*4/*4
CYP2D6_4.021_CYP2D6_4.021	*4/*4	*4	*4	*4,*10,*39	*4,*10,*39	*4/*4	*4/*4	*4/*4	*4/*4
CYP2D6_4.026_CYP2D6_4.026	*4/*4	*4	*4	*4,*10,*39	*4,*10,*39	*4/*4	*4/*4	*4/*4	*4/*4
CYP2D6_6.001_CYP2D6_6.001	*6/*6	*6	*6	*6	*6	*6/*6	*6/*6	*6/*6	*6/*6
CYP2D6_6.002_CYP2D6_6.002	*6/*6	*6	*6	*6	*6	*6/*6	*6/*6	*6/*6	*6/*6
CYP2D6_6.003_CYP2D6_6.003	*6/*6	*6	*6	*6	*6	*6/*6	*6/*6	*6/*6	*6/*6
CYP2D6_6.004_CYP2D6_6.004	*6/*6	*6	*6	*6	*6	*6/*6	*6/*6	*6/*6	*6/*6
CYP2D6_6.005_CYP2D6_6.005	*6/*6	*6	*6	*6	*6	*6/*6	*6/*6	*6/*6	*6/*6
CYP2D6_6.006_CYP2D6_6.006	*6/*6	*6	*6	*6	*6	*6/*6	*6/*6	*6/*6	*6/*6
CYP2D6_7.001_CYP2D6_7.001	*7/*7	*7	*7	*7	*7	*7/*7	*7/*7	*7/*7	*7/*7
CYP2D6_8.001_CYP2D6_8.001	*8/*8	*8	*8	*8,*2,*2D,*39,*34	*8,*2,*2D,*39,*34	*8/*8	*8/*8	*8/*8	*8/*8
CYP2D6_9.001_CYP2D6_9.001	*9/*9	*9	*9	*9	*9	*9/*9	*9/*9	*9/*9	*9/*9
CYP2D6_9.002_CYP2D6_9.002	*9/*9	*9	*9	*9	*9	*9/*9	*9/*9	*9/*9	*9/*9
CYP2D6_10.001_CYP2D6_10.001	*10/*10	*10	*10	*10,*39	*10,*39	*10/*10	*10/*10	*10/*10	*10/*10
CYP2D6_10.002_CYP2D6_10.002	*10/*10	*10	*10	*10,*39	*10,*39	*10/*10	*10/*10	*10/*10	*10/*10
CYP2D6_10.004_CYP2D6_10.004	*10/*10	*10	*10	*10,*39	*10,*39	*10/*10	*10/*10	*10/*10	*10/*10
CYP2D6_11.001_CYP2D6_11.001	*11/*11	*11	*11	*11,*2,*2D,*39,*34	*11,*2,*2D,*39,*34	*11/*11	*11/*11	*11/*11	*11/*11
CYP2D6_12.001_CYP2D6_12.001	*12/*12	*12	*12	*12,*2,*2D,*39,*34	*12,*2,*2D,*39,*34	*12/*12	*12/*12	*12/*12	*12/*12
CYP2D6_12.002_CYP2D6_12.002	*12/*12	*12	*12	*12,*45,*2D,*34	*12,*45,*2D,*34	*12/*12	*12/*12	*12/*45	*12/*12
CYP2D6_14.001_CYP2D6_14.001	*14/*14	*14	*14	*14,*2,*2D,*39,*34	*14,*2,*2D,*39,*34	*14/*14	*14/*14	*14/*14	*14/*14
CYP2D6_15.001_CYP2D6_15.001	*15/*15	*15	*15	*15	*15	*15/*15	*15/*15	*15/*15	*15/*15
CYP2D6_15.002_CYP2D6_15.002	*15/*15	*15	*15	*15	*15	*15/*15	*15/*15	*15/*15	*15/*15

Table S3.1 – continued

Test Case	Truth diplotype	Stargazer				Astrolabe		Aldy	
		hap1_main	hap2_main	hap1_candidate	hap2_canditate	Run 1	Run 2	Run 1	Run 2
CYP2D6_15.003_CYP2D6_15.003	*15/*15	*15	*15	*15,*43	*15,*43	[mac]	[mac]	NG	NG
CYP2D6_17.001_CYP2D6_17.001	*17/*17	*17	*17	*17,*2,*2D,*39,*34	*17,*2,*2D,*39,*34	*17/*17	*17/*17	*17/*17	*17/*17
CYP2D6_17.002_CYP2D6_17.002	*17/*17	*17	*17	*17,*2,*2D,*39,*34	*17,*2,*2D,*39,*34	*17/*17	*17/*17	*17/*17	*17/*17
CYP2D6_17.003_CYP2D6_17.003	*17/*17	*17	*17	*17,*2,*2D,*39,*34	*17,*2,*2D,*39,*34	*17/*17	*17/*17	*17/*17	*17/*17
CYP2D6_18.001_CYP2D6_18.001	*18/*18	*18	*18	*18	*18	*1/*1	*1/*1	*1/*1	*1/*1
CYP2D6_19.001_CYP2D6_19.001	*19/*19	*19	*19	*19,*2,*2D,*39,*34	*19,*2,*2D,*39,*34	*19/*19	*19/*19	*19/*19	*19/*19
CYP2D6_20.001_CYP2D6_20.001	*20/*20	*2	*2	*2,*2D,*39,*34	*2,*2D,*39,*34	*20/*20	*20/*20	*2/*2	*20/*20
CYP2D6_21.001_CYP2D6_21.001	*21/*21	*21	*21	*21,*2,*2D,*39,*34	*21,*2,*2D,*39,*34	*21/*21	*21/*21	*21/*21	*21/*21
CYP2D6_21.002_CYP2D6_21.002	*21/*21	*21	*21	*21,*2,*2D,*39,*34	*21,*2,*2D,*39,*34	*21/*21	*21/*21	*21/*21	*21/*21
CYP2D6_22.001_CYP2D6_22.001	*22/*22	*22	*22	*22	*22	*22/*22	*22/*22	*22/*22	*22/*22
CYP2D6_23.001_CYP2D6_23.001	*23/*23	*23	*23	*23	*23	*23/*23	*23/*23	*23/*23	*23/*23
CYP2D6_24.001_CYP2D6_24.001	*24/*24	*24	*24	*24	*24	*24/*24	*24/*24	*24/*24	*24/*24
CYP2D6_25.001_CYP2D6_25.001	*25/*25	*25	*25	*25	*25	*25/*25	*25/*25	*25/*25	*25/*25
CYP2D6_26.001_CYP2D6_26.001	*26/*26	*26	*26	*26	*26	*26/*26	*26/*26	*26/*26	*26/*26
CYP2D6_27.001_CYP2D6_27.001	*27/*27	*27	*27	*27	*27	*27/*27	*27/*27	*27/*27	*27/*27
CYP2D6_28.001_CYP2D6_28.001	*28/*28	*28	*28	*28,*2,*2D,*39,*34	*28,*2,*2D,*39,*34	*28/*28	*28/*28	*28/*28	*28/*28
CYP2D6_28.002_CYP2D6_28.002	*28/*28	*28	*28	*28,*2,*2D,*39,*34	*28,*2,*2D,*39,*34	*2/*2	*28/*28	*28/*28	*28/*28
CYP2D6_29.001_CYP2D6_29.001	*29/*29	*29	*29	*29,*2,*2D,*39,*34,*107	*29,*2,*2D,*39,*34,*107	*29/*29	*29/*29	*29/*29	*29/*29
CYP2D6_30.001_CYP2D6_30.001	*30/*30	*2	*2	*2,*2D,*39,*30,*34	*2,*2D,*39,*30,*34	*30/*30	*30/*30	*30/*30	*30/*30
CYP2D6_31.001_CYP2D6_31.001	*31/*31	*31	*31	*31,*2,*2D,*39,*34	*31,*2,*2D,*39,*34	*31/*31	*31/*31	*31/*31	*31/*31
CYP2D6_33.001_CYP2D6_33.001	*33/*33	*33	*33	*33	*33	*33/*33	*33/*33	*33/*33	*33/*33
CYP2D6_34.001_CYP2D6_34.001	*34/*34	*34	*34	*34	*34	*34/*34	*34/*34	*34/*34	*34/*34
CYP2D6_35.001_CYP2D6_35.001	*35/*35	*35	*35	*35,*2,*2D,*39,*34	*35,*2,*2D,*39,*34	*35/*35	*35/*35	*35/*35	*35/*35
CYP2D6_35.002_CYP2D6_35.002	*35/*35	*35	*35	*35,*2,*2D,*39,*34	*35,*2,*2D,*39,*34	*35/*35	*35/*35	*35/*35	*35/*35
CYP2D6_35.003_CYP2D6_35.003	*35/*35	*35	*35	*35,*2,*2D,*39,*34	*35,*2,*2D,*39,*34	*35/*35	*35/*35	*35/*35	*35/*35
CYP2D6_35.004_CYP2D6_35.004	*35/*35	*35	*35	*35,*2,*2D,*39,*34	*35,*2,*2D,*39,*34	*35/*35	*35/*35	*35/*35	*35/*35
CYP2D6_35.006_CYP2D6_35.006	*35/*35	*35	*35	*35,*2,*2D,*39,*34	*35,*2,*2D,*39,*34	*35/*35	*35/*35	*35/*35	*35/*35
CYP2D6_35.007_CYP2D6_35.007	*35/*35	*35	*35	*35,*2,*2D,*39,*34	*35,*2,*2D,*39,*34	*35/*35	*35/*35	*35/*35	*35/*35
CYP2D6_37.001_CYP2D6_37.001	*37/*37	*10	*10	*10,*37,*39	*10,*37,*39	*37/*37	*37/*37	*37/*37	*37/*37
CYP2D6_38.001_CYP2D6_38.001	*38/*38	*38	*38	*38	*38	*38/*38	*38/*38	*1/*1	*38/*38
CYP2D6_39.001_CYP2D6_39.001	*39/*39	*39	*39	*39	*39	*39/*39	*39/*39	*39/*39	*39/*39
CYP2D6_40.001_CYP2D6_40.001	*40/*40	*40	*40	*40,*17,*2,*2D,*39,*34	*40,*17,*2,*2D,*39,*34	*40/*40	*40/*40	*40/*40	*40/*40
CYP2D6_41.001_CYP2D6_41.001	*41/*41	*41	*41	*41,*2,*2D,*39,*34	*41,*2,*2D,*39,*34	*41/*41	*41/*41	*41/*41	*41/*41
CYP2D6_41.002_CYP2D6_41.002	*41/*41	*41	*41	*41,*119,*2,*2D,*39,*34	*41,*119,*2,*2D,*39,*34	*41/*41	*41/*41	*41/*41	*41/*41
CYP2D6_41.003_CYP2D6_41.003	*41/*41	*41	*41	*41,*119,*2,*2D,*39,*34	*41,*119,*2,*2D,*39,*34	*41/*41	*41/*41	*41/*41	*41/*41
CYP2D6_41.004_CYP2D6_41.004	*41/*41	*41	*41	*41,*119,*2,*2D,*39,*34	*41,*119,*2,*2D,*39,*34	*41/*41	*41/*41	*41/*41	*41/*41
CYP2D6_41.005_CYP2D6_41.005	*41/*41	*41	*41	*41,*119,*2,*2D,*39,*34	*41,*119,*2,*2D,*39,*34	*41/*41	*41/*41	*41/*41	*41/*41
CYP2D6_42.001_CYP2D6_42.001	*42/*42	*42	*42	*42,*2,*2D,*39,*34	*42,*2,*2D,*39,*34	*42/*42	*42/*42	*42/*42	*42/*42
CYP2D6_43.001_CYP2D6_43.001	*43/*43	*43	*43	*43	*43	*43/*43	*43/*43	*43/*43	*43/*43
CYP2D6_43.002_CYP2D6_43.002	*43/*43	*43	*43	*43	*43	*43/*43	*43/*43	*43/*43	*43/*43
CYP2D6_44.001_CYP2D6_44.001	*44/*44	*44	*44	*44,*22	*44,*22	*44/*44	*44/*44	*44/*44	*44/*44
CYP2D6_45.001_CYP2D6_45.001	*45/*45	*2	*2	*2,*45,*2D,*39,*34	*2,*45,*2D,*39,*34	*45/*45	*45/*45	*45/*45	*45/*45
CYP2D6_45.002_CYP2D6_45.002	*45/*45	*2	*2	*2,*45,*2D,*39,*34	*2,*45,*2D,*39,*34	*45/*45	*45/*45	*45/*45	*45/*45
CYP2D6_46.001_CYP2D6_46.001	*46/*46	*46	*46	*46,*2,*45,*2D,*39,*34,*43	*46,*2,*45,*2D,*39,*34,*43	*46/*46	*46/*46	*46/*46	*46/*46
CYP2D6_46.002_CYP2D6_46.002	*46/*46	*46	*46	*46,*2,*45,*2D,*39,*34,*43	*46,*2,*45,*2D,*39,*34,*43	*46/*46	*46/*46	*46/*46	*46/*46
CYP2D6_46.003_CYP2D6_46.003	*46/*46	*46	*46	*46,*2,*45,*2D,*39,*34,*43	*46,*2,*45,*2D,*39,*34,*43	*46/*46	*46/*46	*46/*46	*46/*46
CYP2D6_47.001_CYP2D6_47.001	*47/*47	*47	*47	*47,*10,*39	*47,*10,*39	*47/*47	*47/*47	*47/*47	*47/*47
CYP2D6_48.001_CYP2D6_48.001	*48/*48	*48	*48	*48	*48	*48/*48	*48/*48	*48/*48	*48/*48
CYP2D6_49.001_CYP2D6_49.001	*49/*49	*49	*49	*49,*10,*39	*49,*10,*39	*49/*49	*49/*49	*49/*49	*49/*49
CYP2D6_50.001_CYP2D6_50.001	*50/*50	*50	*50	*50	*50	*50/*50	*50/*50	*50/*50	*50/*50
CYP2D6_51.001_CYP2D6_51.001	*51/*51	*51	*51	*51,*2,*2D,*39,*34	*51,*2,*2D,*39,*34	*51/*51	*51/*51	*51/*51	*51/*51
CYP2D6_52.001_CYP2D6_52.001	*52/*52	*10	*10	*10,*52,*39,*106	*10,*52,*39,*106	*52/*52	*52/*52	*52/*52	*52/*52
CYP2D6_52.002_CYP2D6_52.002	*52/*52	*10	*10	*10,*52,*39,*106	*10,*52,*39,*106	*10/*10	*52/*52	*52/*52	*52/*52
CYP2D6_53.001_CYP2D6_53.001	*53/*53	*53	*53	*53	*53	*53/*53	*53/*53	*53/*53	*53/*53

Table S3.1 – continued

Test Case	Truth diplotype	Stargazer				Astrolabe		Aldy	
		hap1_main	hap2_main	hap1_candidate	hap2_canditate	Run 1	Run 2	Run 1	Run 2
CYP2D6_54.001_CYP2D6_54.001	*54/*54	*54	*54	*54,*10,*39	*54,*10,*39	*54/*54	*54/*54	*54/*54	*54/*54
CYP2D6_55.001_CYP2D6_55.001	*55/*55	*55	*55	*55,*2,*2D,*39,*34	*55,*2,*2D,*39,*34	*55/*55	*55/*55	*55/*55	*55/*55
CYP2D6_56.001_CYP2D6_56.001	*56/*56	*56	*56	*56,*2,*2D,*39,*34	*56,*2,*2D,*39,*34	*56/*56	*56/*56	*56/*56	*56/*56
CYP2D6_56.002_CYP2D6_56.002	*56/*56	*56	*56	*56,*10,*39	*56,*10,*39	*56/*56	*56/*56	*56/*56	*56/*56
CYP2D6_56.003_CYP2D6_56.003	*56/*56	*56	*56	*56,*10,*39	*56,*10,*39	[mac]	*56/*56	*56/*56	*56/*56
CYP2D6_58.001_CYP2D6_58.001	*58/*58	*17	*17	*17,*58,*2,*2D,*39,*30,*34	*17,*58,*2,*2D,*39,*30,*34	*58/*58	*58/*58	*58/*58	*58/*58
CYP2D6_59.001_CYP2D6_59.001	*59/*59	*59	*59	*59,*2,*2D,*39,*34	*59,*2,*2D,*39,*34	*59/*59	*59/*59	*59/*59	*59/*59
CYP2D6_60.001_CYP2D6_60.001	*60/*60	*60	*60	*60	*60	*60/*60	*60/*60	*60/*60	*60/*60
CYP2D6_62.001_CYP2D6_62.001	*62/*62	*62	*62	*62	*62	*62/*62	*62/*62	*62/*62	*62/*62
CYP2D6_64.001_CYP2D6_64.001	*64/*64	*10	*10	*10,*17,*64,*39	*10,*17,*64,*39	*64/*64	*64/*64	*64/*64	*64/*64
CYP2D6_65.001_CYP2D6_65.001	*65/*65	*10	*10	*10,*65,*2,*2D,*39,*34	*10,*65,*2,*2D,*39,*34	*65/*65	*65/*65	*65/*65	*65/*65
CYP2D6_69.001_CYP2D6_69.001	*69/*69	*69	*69	*69,*10,*41,*65,*119,*2,*2D,*39,*34	*69,*10,*41,*65,*119,*2,*2D,*39,*34	*69/*69	*69/*69	*69/*69	*69/*69
CYP2D6_70.001_CYP2D6_70.001	*70/*70	*70	*70	*70,*39,*107	*70,*39,*107	*70/*70	*70/*70	*70/*70	*70/*70
CYP2D6_71.001_CYP2D6_71.001	*71/*71	*71	*71	*71	*71	*71/*71	*71/*71	*71/*71	*71/*71
CYP2D6_71.002_CYP2D6_71.002	*71/*71	*71	*71	*71	*71	*71/*71	*71/*71	*71/*71	*71/*71
CYP2D6_71.003_CYP2D6_71.003	*71/*71	*71	*71	*71	*71	*71/*71	*71/*71	[mac]	*71/*71
CYP2D6_72.001_CYP2D6_72.001	*72/*72	*72	*72	*72,*10,*39	*72,*10,*39	*72/*72	*72/*72	*72/*72	*72/*72
CYP2D6_73.001_CYP2D6_73.001	*73/*73	*2	*2	*2,*2D,*39,*34,*73	*2,*2D,*39,*34,*73	*73/*73	*73/*73	*73/*73	*73/*73
CYP2D6_74.001_CYP2D6_74.001	*74/*74	*74	*74	*74	*74	*74/*74	*74/*74	*74/*74	*74/*74
CYP2D6_75.001_CYP2D6_75.001	*75/*75	*75	*75	*75	*75	*75/*75	*75/*75	*75/*75	*75/*75
CYP2D6_81.001_CYP2D6_81.001	*81/*81	*81	*81	*81,*86	*81,*86	*81/*81	*81/*81	*81/*81	*81/*81
CYP2D6_84.001_CYP2D6_84.001	*84/*84	*84	*84	*84,*2,*2D,*39,*34	*84,*2,*2D,*39,*34	*84/*84	*84/*84	*84/*84	*84/*84
CYP2D6_84.002_CYP2D6_84.002	*84/*84	*84	*84	*84,*2,*2D,*39,*34	*84,*2,*2D,*39,*34	*84/*84	*84/*84	*84/*84	*84/*84
CYP2D6_85.001_CYP2D6_85.001	*85/*85	*85	*85	*85,*2,*2D,*39,*34	*85,*2,*2D,*39,*34	*85/*85	*85/*85	*85/*85	*85/*85
CYP2D6_86.001_CYP2D6_86.001	*86/*86	*86	*86	*86	*86	*86/*86	*86/*86	*86/*86	*86/*86
CYP2D6_87.001_CYP2D6_87.001	*87/*87	*10	*10	*10,*87,*39	*10,*87,*39	*87/*87	*87/*87	*87/*87	*87/*87
CYP2D6_88.001_CYP2D6_88.001	*88/*88	*88	*88	*88,*39	*88,*39	*88/*88	*88/*88	*88/*88	*88/*88
CYP2D6_89.001_CYP2D6_89.001	*89/*89	*89	*89	*89	*89	*89/*89	*89/*89	*89/*89	*89/*89
CYP2D6_90.001_CYP2D6_90.001	*90/*90	*90	*90	*90	*90	*90/*90	*90/*90	*90/*90	*90/*90
CYP2D6_91.001_CYP2D6_91.001	*91/*91	*41	*41	*41,*91,*34	*41,*91,*34	*91/*91	*91/*91	*91/*91	*91/*91
CYP2D6_92.001_CYP2D6_92.001	*92/*92	*92	*92	*92	*92	*92/*92	*92/*92	*92/*92	*92/*92
CYP2D6_93.001_CYP2D6_93.001	*93/*93	*93	*93	*93	*93	*93/*93	*93/*93	*93/*93	*93/*93
CYP2D6_94.001_CYP2D6_94.001	*94/*94	*10	*10	*10,*94,*39	*10,*94,*39	*94/*94	*94/*94	*94/*94	*94/*94
CYP2D6_94.002_CYP2D6_94.002	*94/*94	*10	*10	*10,*94,*39	*10,*94,*39	*94/*94	*94/*94	*94/*94	*94/*94
CYP2D6_95.001_CYP2D6_95.001	*95/*95	*10	*10	*10,*95,*39	*10,*95,*39	*95/*95	*95/*95	*95/*95	*95/*95
CYP2D6_96.001_CYP2D6_96.001	*96/*96	*96	*96	*96	*96	*96/*96	*96/*96	*96/*96	*96/*96
CYP2D6_97.001_CYP2D6_97.001	*97/*97	*97	*97	*97	*97	*97/*97	*97/*97	*97/*97	*97/*97
CYP2D6_98.001_CYP2D6_98.001	*98/*98	*2	*2	*2,*2D,*39,*34,*98	*2,*2D,*39,*34,*98	*98/*98	*98/*98	*98/*98	*98/*98
CYP2D6_100.001_CYP2D6_100.001	*100/*100	*100	*100	*100,*10,*39	*100,*10,*39	*100/*100	*100/*100	*100/*100	*100/*100
CYP2D6_101.001_CYP2D6_101.001	*101/*101	*101	*101	*101,*10,*39	*101,*10,*39	*101/*101	*101/*101	*101/*101	*101/*101
CYP2D6_102.001_CYP2D6_102.001	*102/*102	*2	*2	*2,*102,*2D,*39,*34,*48	*2,*102,*2D,*39,*34,*48	*102/*102	*102/*102	*102/*102	*102/*102
CYP2D6_103.001_CYP2D6_103.001	*103/*103	*103	*103	*103,*2,*102,*2D,*39,*34,*48	*103,*2,*102,*2D,*39,*34,*48	*103/*103	*103/*103	*103/*103	*103/*103
CYP2D6_104.001_CYP2D6_104.001	*104/*104	*104	*104	*104,*2,*2D,*39,*34	*104,*2,*2D,*39,*34	*104/*104	*104/*104	*104/*104	*104/*104
CYP2D6_105.001_CYP2D6_105.001	*105/*105	*105	*105	*105,*2,*2D,*39,*34	*105,*2,*2D,*39,*34	*105/*105	*105/*105	*105/*105	*105/*105
CYP2D6_106.001_CYP2D6_106.001	*106/*106	*106	*106	*106	*106	*1/*1	*106/*106	*106/*106	*106/*106
CYP2D6_106.002_CYP2D6_106.002	*106/*106	*106	*106	*106	*106	*1/*1	*106/*106	*106/*106	*106/*106
CYP2D6_107.001_CYP2D6_107.001	*107/*107	*107	*107	*107	*107	*107/*107	*107/*107	*107/*107	*107/*107
CYP2D6_108.001_CYP2D6_108.001	*108/*108	*108	*108	*108	*108	*108/*108	*108/*108	*108/*108	*108/*108
CYP2D6_109.001_CYP2D6_109.001	*109/*109	*9	*9	*9,*109	*9,*109	*109/*109	*109/*109	*109/*109	*109/*109

Run1/Run2 — For Astrolabe and Aldy, run 1 and run 2 indicate results before and after updating their allele definitions respectively; g, genotyped; NG, not genotyped due to error message; [mac], ambiguous call; For Stargazer, cases where correct haplotypes are reported under hap1_candidate or hap2_candidate but not under hap1_main or hap2_main are coloured blue (N=18)

Table S3.2: Main allele calls for simulated diplotype combinations (starting from BAM files) of haplotypes with definitive or moderate PharmVar level of evidence. This table is too large to be reprinted in this Thesis. The full table can be found at
https://static-content.springer.com/esm/art%3A10.1038%2Fs41525-020-0135-2/MediaObjects/41525_2020_135_MOESM4_ESM.xlsx

Table S3.3: Astrolabe, Aldy, and Stargazer calls for samples with simulated *CYP2D6* structural variants and 30x read depth

Read depth: 30x										
Sample ID	**Truth**	**Aldy**			**Astrolabe**			**Stargazer**		
		Hap1	**Hap2**	**Notes**	**Hap1**	**Hap2**	**Notes**	**Hap1**	**Hap2**	**Notes**
SIM101	**79+*68A+*4.009/*3.001*	**3-like*	**4+*4*		**3*	**4*		**3*	**4x2*	
SIM102	**1.002X2+*83.001/*56.002*	**1+*1*	**56+*83*		**1*	**56*	*(+)*	**1x2*	**56x2*	
SIM103	**1.001+*90/*27.001*	**1+*90*	**27*		**27*	**90*		**1*	**1x2*	
SIM104	**10.001X2/*4.001*	**10+*10*	**4*		**10*	**4*	*(+)*	**4*	**10x2*	
SIM105	**10.002X3/*42.001*	**10+*10+*10*	**42*		**10*	**42*	*(+)*	**10x3*	**42*	
SIM106	**76+*1.002/*5*	**1*	**76*		**1 [*27]*	**5*		**1*	**13C*	
SIM107	**79+*68A+*4.001x2/*7.001*	**4+*7*	**4-like+*80*		**7*	**4*		**4x2*	**7x2*	
SIM108	**17X2(REPdup)/*43.001 x2*	**17+*17*	**43+*43*		**17*	**43*	*(+)*	**17x2*	**43x2*	
SIM109	**17X2(spacer)/*69.001*	**17+*83*	**69*		**17*	**69*		**2*	**83+*2*	
SIM110	**1.002X3/*59.001*	**1+*1+*1*	**59*		**1*	**59*	*(+)*	**1x3*	**59*	
SIM111	**1.004X5/*4.001*	**1+*1+*1+*1*	**1-like*		**1*	**4*	*(+)*	**1x5*	**113*	
SIM112	**80/*5*	**5*	**67*		*{*2M}*	*{*2M}*		**5*	**13B*	
SIM113	**29X2/*84.001*	**29+*29*	**84*		**29*	**84*	*(+)*	**29x2*	**84*	
SIM114	**2.007X2/*70.001*	**2+*2*	**70*		**2*	**70*	*(+)*	**2x2*	**39*	
SIM115	**2.010X4/*58.001*	**2+*2+*2*	**58*		**2*	**58*	*(+)*	**2x2*	**17x3*	
SIM116	**2.001x8/*29.001x2*	**2+*2+*2+*2+*2+*2*	**29+*63+*80*		**2*	**29*	*(+)*	**2x4*	**29x6*	
SIM117	**35.001X2/*95.001*	**35+*35*	**95*		**35*	**95*	*(+)*	**10*	**35x2*	
SIM118	**36.001(REP6)/*4.013*	**36*	**4*		**5*	**4*		**4*	**10*	
SIM119	**36.001+*10.001/*36.001+*10.001*	**36+*10*	**36+*10*		**10*	**10*	*(+)*	**36+*10*	**36+*10*	
SIM120	**36.001X2/*10.002*	**10*	**36+*10*		**10*	**10*		**10*	**36+*10*	
SIM121	**36.001X2+*10.003/*107.001*	**107+*36*	**36+*10*		**10*	**107*		.	.	
SIM122	**36.002/*3.001X2*	**3+*3*	**36*		**3*	**10*		**1*	**3x2*	
SIM123	**36+*10.001x2/*64.001*	**10+*64*	**36+*10*		**10*	**64*		**10x2*	**36+*10*	
SIM124	**83.001/*54.001*	**54*	**61*		**1*	**54*		**1*	**10*	
SIM125	**4.013+*4.001/*12.002*	**12+*4N*	**45*		**4*	**12*		**4N+*4*	**12*	
SIM126	**4.013X2/*10.004*	**36*	**4+*4*		**10*	**4*		**4N+*4*	**10*	
SIM127	**41.001X2/*68+*4.001*	**2+*41*	**68+*4*		**41*	**4*		**S1+*1*	**1x2*	
SIM128	**68/*4.001*	**4*	**68-like*		**5*	**4*		**S1+*1*	**13C*	
SIM129	**43.001X2/*60.001*	**43+*43*	**60*		**43*	**60*	*(+)*	**43x2*	**60*	
SIM130	**45X2/*62.001*	**45+*45*	**62*		**62*	**45*		**1x2*	**62*	
SIM131	**4.011X2/*16*	**16*	**4+*4*		**10*	**4*		**4*	**4*	
SIM132	**4.001X3/*83.003*	**36+*80*	**4*		**1*	**4*		**4x2*	**39x2*	
SIM133	**5/*5*	**5*	**5*		**2M*	**2M*		**5*	**5*	
SIM134	**57+*10D/*45.001*	**45*	**57+*10*		**10*	**45*		**10*	**83+*2*	
SIM135	**61/*4.008*	**4*	**61*		**5*	**4*		**4N+*4*	**5*	
SIM136	**63/*15.001*	**15*	**63*		**2*	**15*		*NG*	*NG*	
SIM137	**13A1/*11.001*	**1*	**79*		**1*	**11*		**11*	**13B*	
SIM138	**67/*9.001X2*	**1 +*9*	**67*		**1*	**9*		*NG*	*NG*	
SIM139	**68+*4.001/*18.001*	**1*	**4+*68-like*		**18*	**4*		**S1+*1*	**4*	
SIM140	**6.003X2/*71.001*	**6+*6*	**71*		**6*	**71*	*(+)*	**6x2*	**71*	
SIM141	**13A1+*1.004x2/*43.001*	**1+*1*	**43+*80*		**1*	**43*	*(+)*	*NG*	*NG*	
SIM142	**57+*10D/*35.002*	**35*	**57+*10*		**10*	**35*	*(+)*	**10*	**83+*2*	

Table S3.3 – continued

Sample ID	Truth	Aldy			Astrolabe			Stargazer		
		Hap1	Hap2	Notes	Hap1	Hap2	Notes	Hap1	Hap2	Notes
SIM143	*77+*2.002/*2.011	*2	*79+*2		*2	*2		NG	NG	
SIM144	*78+*2.001/*34.001	*2+*67	*34		*2	*34	(+)	*2	*S2+*1	
SIM145	*82.001/*33.001	*1-like	*33		*33	*82		*1	*33	
SIM146	*79+*2.001/*39.001	*39	*79+*2		*2	*4		NG	NG	
SIM147	*83.002/*5	*1	*5		*1	*5		*1	*5	
SIM148	*80/*4.004	*4	*67		*4	*4		*4	*13C	
SIM149	*29x2/*5	*29	*29		*29	*29		*29	*29	
SIM150	*1.002x4/*2.001x8	*1+*1+*1+*63	*2x7+*79		*1	*2	(+)	*1x6	*34x6	
SIM167	*4/*68+*4	*4+*4	*68-like		*4	*4		*S1+*1	*4	
SIM168	*68+*4/*68+*4	*4	*68-like		*4	*4		*S1+*1	*S1+*1	
SIM169	*79+*68+*4.009/*2.011	*2	*4+*4		*2	*4		*4x2	*34	
SIM170	*1.002x2+*83.001/*6.001	*1+*1	*6+*61		*1	*6	(+)	*1x2	*6x2	
SIM171	*1.001+*90.001/*4.001	*1+*90	*4		*90	*4		*1x2	*4	
SIM172	*10.001x2/*3.001	*10+*10	*3		*3	*10	(+)	*3	*10x2	
SIM173	*3.001x2/*4.001x2	*3+*3	*4+*4		*3	*4	(+)	*1x2	*3x2	
SIM174	*76+*1.002/*34.001	*34	*76+*1		*1	*27		*S2+*1	*34	
SIM175	*79+*68+*4.001x2/*40.001	*2+*40	*4+*4		*4	*40		*4x3	*40	
SIM176	*17.001x2/*29.001x2	*17+*17	*29+*29		*17	*29	(+)	*17x2	*29x2	
SIM177	*17x2(spacer)/*45.001	*17+*61	*45		*17	*45	(+)	*2	*83+*2	
SIM178	*1.002x3/*2.001x2	*1+*1+*1	*2+*2		*1	*2	(+)	*1x3	*2x2	
SIM179	*80/*33.001	*33	*78		*1	*33		*13B	*33	
SIM180	*29.001x2/*4.004	*29+*29	*4		*29	*4		*4	*29x2	
SIM181	*35.001x2/*1.002	*1	*35+*35		*1	*35		*1	*35x2	
SIM182	*36.001/*9.001	*10	*9		*9	*10		*1	*9	
SIM183	*36.001x2/*100.001	*100	*36+*10		*10	*100		*36+*10	*100	
SIM184	*4.013+*4.001/*50.001	*4+*4N	*50		*4	*50		*4N+*4	*50	
SIM185	*4.013x2/*51.001	*4+*4N	*51		*4	*51		*4N+*4	*10	
SIM186	*43.001x2/*28.001	*28	*43+*43		*28	*43	(+)	*1x2	*2D	
SIM187	*5/*5	*5	*5		*2	*2		*5	*5	
SIM188	*27.001/*5	*27	*5		*27	*5		*1	*5	
SIM189	*57.001+*10.003/*1.001	*1	*57+*10		*1	*10	(+)	NG	NG	
SIM190	*61/*84.001	*61	*84		*1	*84		*5	*83+*2	
SIM191	*63/*58.001	*58	*63		*2	*58		*5	*83+*2	
SIM192	*61/*1.001x3	*1+*1+*1	*61		*1	*1	(+)	NG	NG	
SIM193	*63/*2.001x2	*2+*2	*63		*2	*2		*2	*83+*2	
SIM194	*13A1/*9x2	*1	*79		*1	*9		*9	*13B	
SIM195	*67/*81	*1	*67		*81	*1		*13C	*81	
SIM196	13A+*1.004x2/*7	*1+*1	*7+*79		*1	*7	(+)	NG	NG	
SIM197	*77+*2.002/*3.001	*3	*79+*2		*2	*3		NG	NG	
SIM198	*78+*2.001/*39.001	*2+*67	*39		*2	*39	(+)	*2	*S2+*1	
SIM199	*83.002/*2.001	*2	*39		*1	*2		*1	*2	
SIM200	*82.001/*2.011	*2	*82		*1	*2		*1	*2	
SIM201	*2.001x8/*1.001	*2+*2+*2+*2+*2+*2	*61+*79+*2		*2	*34	(+)	*1	*2x8	
SIM202	*1.001x4/*2.001	*1+*1+*1+*1	*2		*2	*2	(+)	*1x2	*2Dx3	
SIM203	*2.011/*2.011	*2	*2		*2	*2		*2	*2	
SIM277	*13A+*1.004x2/*13A+*1.004x2	*1+*1+*1+*1	*79+*79		*1	*1		NG	NG	
SIM278	*13A/*13A	*13	*79		*1	*1		*13B	*13B	
SIM279	*1.001+*90.001/*1.001+*90.001	*1+*1	*90+*90		*1	*90		*1	*1x3	

Table S3.3 – continued

Sample ID	Truth	Aldy			Astrolabe			Stargazer		
		Hap1	Hap2	Notes	Hap1	Hap2	Notes	Hap1	Hap2	Notes
SIM280	**1.002X2+*83.001/*1.002x2+*83.001*	**1+*1+*1+*1*	**61*		**1*	**1*	*(+)*	**1*	**1x5*	
SIM281	**36.001/*36.001*	**10*	**36*		**10*	**10*		**10*	**10*	
SIM282	**36.001x2/*36.001x2*	**36+*10*	**36+*10*		**10*	**10*		**36+*10*	**36+*10*	
SIM283	**36.001x2+*10.003/*36x2+*10.003*	**36+*10*	**10+*10+*36+*10*		**10*	**10*	*(+)*	*NG*	*NG*	
SIM284	**61/*61*	*NG*	*NG*		**1*	**5*		*NG*	*NG*	
SIM285	**63/*63*	**63*	**63*		**2*	**2 [*5]*		**5*	**83+*2*	
SIM286	**57+*10D/*57+*10D*	**57+*10*	**57+*10*		**10*	**10*		*NG*	*NG*	
SIM287	**76+*1.002/*76+1.002*	**1+*1*	**77*		**1*	**27*		*NG*	*NG*	
SIM288	**77+*2.002/*77+*2.002*	**13+*2*	**79+*2*		**2*	**2*		*NG*	*NG*	
SIM289	**78+*2.001/*78+*2.002*	**2+*67*	**78+*2*		**2*	**2*	*(+)*	*NG*	*NG*	
SIM290	**79+*68+*4.001x2/*79+*68+*4.001x2*	**4+*4+*4+*4+*4*	**4-like*		**4*	**4*		**4*	**4x5*	
SIM291	**80/*80*	**67*	**78*		**2*	**2*		**13C*	**13B*	
SIM292	**82.001/*82.001*	**82*	**82*		**82*	**82*		**1*	**1*	
SIM293	**10.001/*5*	**10*	**5*		**10*	**10*		**5*	**10*	
SIM294	**17.001/*5*	**17*	**5*		**17*	**17*		**5*	**17*	
SIM295	**101.001/*5*	**101*	**5*		**101*	**101*		**5*	**101*	
SIM296	**65.001/*5*	**5*	**65*		**65*	**65*		**5*	**10*	
SIM297	**24.001/*5*	**24*	**5*		**24*	**24*		**1*	**5*	
SIM298	**3.001/*5*	**3*	**5*		**3*	**3*		**3*	**5*	
SIM299	**84.001/*5*	**5*	**84*		**84*	**84*		**5*	**84*	
SIM300	**86.001/*5*	**5*	**86*		**86*	**86*		**5*	**86*	
SIM301	**10.002x3/*5*	**10*	**10+*10*		**10*	**10*		*NG*	*NG*	
SIM302	**79+*68+*4.009/*5*	**4*	**4-like*		**4*	**4*		**4*	**4*	
SIM303	**41x3/*70*	**41+*41+*41*	**70*		**41*	**70*		**2x2*	**41x2*	
SIM304	**2.001x2/*24.001*	**2+*2*	**24*		**2*	**24*		**1*	**2x2*	
SIM305	**35.001x2/*19.001*	**19*	**35+*35*		**19*	**35*		**19*	**35x2*	
SIM306	**43.001x2/*55.001*	**43+*43*	**55*		**43*	**55*		**43x2*	**55*	
SIM307	**29.001x2/*89.001*	**29+*29*	**89*		**29*	**89*		**1*	**29x2*	
SIM308	**17.001x2/*40.001*	**17*	**40+*40*		**17*	**40*		**17x2*	**40*	
SIM309	**10.002x3/*104.001*	**10+*10+*10*	**104*		**10*	**104*		**2*	**10x3*	
SIM310	**10.001x2/*8.001*	**10+*10*	**8*		**8*	**10*		**8*	**10x2*	
SIM311	**79+*68+*4.009/*79+*68+*4.009*	**4+*4+*4-like*	**68+*4+*80*		**4*	**4*	*(+)*	**4*	**4x3*	

NG, not genotyped by tool due to fail error; (+), possible duplication (Astrolabe); , indicates that the call was obtained by running algorithm without BAM quality control due to previous fail; **S1+*1*, Stargazer reports this allele in cases where it does not detect all variants that it uses to define the *CYP2D6*68+*4* tandem rearrangement

Table S3.4: Astrolabe, Aldy, and Stargazer calls for samples with simulated *CYP2D6* structural variants and 60x read depth

Read depth: 60x										
Sample ID	**Truth**	**Aldy**			**Astrolabe**			**Stargazer**		
		Hap1	**Hap2**	**Notes**	**Hap1**	**Hap2**	**Notes**	**Hap1**	**Hap2**	**Notes**
SIM101	**79+*68A+*4.009/*3.001*	**3*	**4+*4-like*		**3*	**4*		**3*	**10x2*	
SIM102	**1.002X2+*83.001/*56.002*	**13+*1*	**61*		**1*	**56*	*(+)*	**1x3*	**56*	
SIM103	**1.001+*90/*27.001*	**1+*90*	**27*		**27*	**90*		**1*	**1x2*	
SIM104	**10.001X2/*4.001*	**10+*10*	**4*		**10*	**4*		**4*	**10x2*	
SIM105	**10.002X3/*42.001*	**10+*10+*10*	**42*		**10*	**42*	*(+)*	**10x3*	**42*	
SIM106	**76+*1.002/*5*	**1*	**76*		**1*	**5*		**1*	**13C*	
SIM107	**79+*68A+*4.001x2/*7.001*	**4+*4+*4*	**7-like*		**7*	**4*	*(+)*	**4x2*	**7x2*	
SIM108	**17X2(REPdup)/*43.001 x2*	**17+*43*	**34*		**17*	**43*		**17x2*	**43x2*	
SIM109	**17X2(spacer)/*69.001*	**17+*83*	**69*		**17*	**69*		**69*	**83+*2*	
SIM110	**1.002X3/*59.001*	**1+*1+*1*	**59*		**1*	**59*	*(+)*	**1x3*	**59*	
SIM111	**1.004X5/*4.001*	**1+*1+*1+*1*	**1-like*		**1*	**113*	*(+)*	**1x5*	**113*	
SIM112	**80/*5*	*NG*	*NG*		**5*	**53*		**5*	**13B*	
SIM113	**29X2/*84.001*	**29+*29*	**84*		**29*	**84*		**29x2*	**84*	
SIM114	**2.007X2/*70.001*	**2+*2*	**70*		**2*	**70*		**2x2*	**39*	
SIM115	**2.010X4/*58.001*	**2+*2+*2*	**58*		**2*	**58*	*(+)*	**2x2*	**17x3*	
SIM116	**2.001x8/*29.001x2*	**29+*29+*63*	**2+*2+*2+*79+*2*		**2*	**29*	*(+)*	**2*	**2x9*	
SIM117	**35.001X2/*95.001*	**35+*35*	**95*		**35*	**95*		**2x2*	**10*	
SIM118	**36.001(REP6)/*4.013*	**36*	**4*		**5*	**4*		**4*	**10*	
SIM119	**36.001+*10.001/*36.001+*10.001*	**36+*10*	**36+*10*		**10*	**10*	*(+)*	**36+*10*	**36+*10*	
SIM120	**36.001X2/*10.002*	**10*	**36+*10*		**10*	**10*		**10*	**36+*10*	
SIM121	**36.001X2+*10.003/*107.001*	**107+*36*	**36+*10*		**10*	**107*		**36+*10*	**36+*10*	
SIM122	**36.002/*3.001X2*	**3+*3*	**36*		**3*	**10*		**3x2*	**10*	
SIM123	**36+*10.001x2/*64.001*	**10+*64*	**36+*10*		**10*	**64*	*(+)*	**10x2*	**36+*10*	
SIM124	**83.001/*54.001*	**39*	**54*		**1*	**54*		**1*	**10*	
SIM125	**4.013+*4.001/*12.002*	**12+*4*	**45*		**12*	**4*		**1*	**4N+*4*	
SIM126	**4.013X2/*10.004*	**36*	**4+*4*		**10*	**4*		**4N+*4*	**10*	
SIM127	**41.001X2/*68+*4.001*	**2+*41*	**68+*4*		**41*	**4*	*(+*68)*	**S1+*1*	**1x2*	
SIM128	**68/*4.001*	**4*	**68-like*		**5*	**4*		**S1+*1*	**13C*	
SIM129	**43.001X2/*60.001*	**43+*43*	**60*		**43*	**60*		**1x2*	**60*	
SIM130	**45X2/*62.001*	**45+*45*	**62*		**62*	**45*		**1x2*	**62*	
SIM131	**4.011X2/*16*	**16*	**4+*4*		**4*	**4*		**4*	**4*	
SIM132	**4.001X3/*83.003*	**13+*61*	**4*		**1*	**4*	*(+)*	**1x2*	**4x2*	
SIM133	**5/*5*	**5*	**5*		*NG*	*NG*		**5*	**5*	
SIM134	**57+*10D/*45.001*	**45*	**57+*10*		**10*	**45*		**10*	**83+*2*	
SIM135	**61/*4.008*	**4*	**61*		**39*	**4*		**4N+*4*	**5*	
SIM136	**63/*15.001*	**15*	**63*		**2*	**15*		*NG*	*NG*	
SIM137	**13A1/*11.001*	**1*	**79*		**1*	**11*		**11*	**13B*	
SIM138	**67/*9.001X2*	**1*	**67*		**1*	**9*		*NG*	*NG*	
SIM139	**68+*4.001/*18.001*	**1-like*	**68+*4*		**1*	**4*	*(+*68)*	**S1+*1*	**4*	
SIM140	**6.003X2/*71.001*	**6+*6*	**71*		**6*	**71*		**6x2*	**71*	
SIM141	**13A1+*1.004x2/*43.001*	**1+*1*	**43+*79*		**1*	**43*		*NG*	*NG*	
SIM142	**57+*10D/*35.002*	**35*	**57+*10*		**10*	**35*		**10*	**83+*2*	

Table S3.4 – continued

Sample ID	Truth	Aldy			Astrolabe			Stargazer		
		Hap1	Hap2	Notes	Hap1	Hap2	Notes	Hap1	Hap2	Notes
SIM143	*77+*2.002/*2.011	*2	*79+*2		*2	*2		NG	NG	
SIM144	*78+*2.001/*34.001	*2+*67	*34		*2	*34		*2	*S2+*1	
SIM145	*82.001/*33.001	*1-like	*33		*33	*82		*1	*33	
SIM146	*79+*2.001/*39.001	*39	*79+*2		*2	*4		NG	NG	
SIM147	*83.002/*5	*39	*5		*1	*5		*1	*5	
SIM148	*80/*4.004	*4	*67		*4	*4		*4	*13B	
SIM149	*29x2/*5	*29	*29		*29	*29		*29	*29	
SIM150	*1.002x4/*2.001x8	*1+*1+*61+*80	*2x8		*1	*2	(+)	*1x6	*2x6	
SIM167	*4/*68+*4	*4+*4	*68-like		*4	*4	(+*68)	*S1+*1	*4	
SIM168	*68+*4/*68+*4	*4+*68-like	*68+*4		*4	*4		*S1+*1	*S1+*1	
SIM169	*79+*68+*4.009/*2.011	*2	*4+*4		*2	*4		*4x2	*10	
SIM170	*1.002x2+*83.001/*6.001	*1+*1	*6+*61		*1	*6	(+)	*1x2	*6x2	
SIM171	*1.001+*90.001/*4.001	*1+*90	*4		*90	*4		*1x2	*4	
SIM172	*10.001x2/*3.001	*10+*10	*3		*3	*10		*3	*10x2	
SIM173	*3.001x2/*4.001x2	*3+*3	*4+*4		*3	*4		*3x2	*4x2	
SIM174	*76+*1.002/*34.001	*34	*76+*1		*1	*27		*S2+*1	*34	
SIM175	*79+*68+*4.001x2/*40.001	*2+*40	*4+*4		*4	*40		*2x2	*4x2	
SIM176	*17.001x2/*29.001x2	*17+*17	*29+*29		*17	*29		*17x2	*29x2	
SIM177	*17x2(spacer)/*45.001	*17+*61	*45		*17	*45		*17	*83+*2	
SIM178	*1.002x3/*2.001x2	*1+*1+*1	*2+*2		*1	*2		*34x2	*39x3	
SIM179	*80/*33.001	*33	*78		*1	*33		*13B	*33	
SIM180	*29.001x2/*4.004	*29+*29	*4		*29	*4		*4	*29x2	
SIM181	*35.001x2/*1.002	*1	*35+*35		*1	*35		*34	*39x2	
SIM182	*36.001/*9.001	*1-like	*9		*9	*10		*9	*10	
SIM183	*36.001x2/*100.001	*100	*36+*10		*100	*10		*36+*10	*100	
SIM184	*4.013+*4.001/*50.001	*4+*4	*50		*50	*4		*4N+*4	*50	
SIM185	*4.013x2/*51.001	*4+*4	*51		*51	*4		*4N+*4	*51	
SIM186	*43.001x2/*28.001	*28	*43+*43		*28	*43		*1x2	*2	
SIM187	*5/*5	*5	*5		.	.		*5	*5	
SIM188	*27.001/*5	*27	*5		*5	*27		*1	*5	
SIM189	*57.001+*10.003/*1.001	*1	*57+*10		*1	*10		NG	NG	
SIM190	*61/*84.001	*61	*84		*1	*84		*5	*83+*2	
SIM191	*63/*58.001	*58	*63		*2	*58		*5	*83+*2	
SIM192	*61/*1.001x3	*1+*1+*1	*61		*1	*1	(+)	NG	NG	
SIM193	*63/*2.001x2	*2+*2	*63		*2	*2		*2	*83+*2	
SIM194	*13A1/*9x2	*1	*79		*1	*9		*9	*13B	
SIM195	*67/*81	*1	*67		*1	*81		*13C	*81	
SIM196	13A+*1.004x2/*7	*1+*1	*7+*79		*1	*7	(+)	NG	NG	
SIM197	*77+*2.002/*3.001	*3	*79+*2		*2	*3		NG	NG	
SIM198	*78+*2.001/*39.001	*2+*67	*39		*2	*39	(+)	*2	*S2+*1	
SIM199	*83.002/*2.001	*2	*61		*1	*2		*2D	*39	
SIM200	*82.001/*2.011	*1-like	*2		*2	*82		*1	*34	
SIM201	*2.001x8/*1.001	*1	*2x8		*1	*2	(+)	*34x4	*39x5	
SIM202	*1.001x4/*2.001	*1+*1+*1	*61+*79		*1	*2	(+)	*1x3	*2x2	
SIM203	*2.011/*2.011	*2	*2		*2	*2		*2	*2	
SIM277	*13A+*1.004x2/*13A+*1.004x2	*1+*1+*1+*1	*79+*79		*1	*1	(+)	NG	NG	
SIM278	*13A/*13A	*13	*79		*1	*1		*13B	*13B	
SIM279	*1.001+*90.001/*1.001+*90.001	*1+*1	*90+*90		*1	*90	(+)	*1	*1x3	

Table S3.4 – continued

Sample ID	Truth	Aldy			Astrolabe			Stargazer		
		Hap1	Hap2	Notes	Hap1	Hap2	Notes	Hap1	Hap2	Notes
SIM280	**1.002X2+*83.001/*1.002x2+*83.001*	**1+*1+*1+*1*	**61+*61*		**1*	**1*	*(+)*	**1*	**1x5*	
SIM281	**36.001/*36.001*	**10*	**36*		**10*	**10*		**10*	**10*	
SIM282	**36.001x2/*36.001x2*	**36+*10*	**36+*10*		**10*	**10*		**36+*10*	**36+*10*	
SIM283	**36.001x2+*10.003/*36x2+*10.003*	**36+*10*	**10+*10+*36+*10*		**10*	**10*	*(+)*	*NG*	*NG*	
SIM284	**61/*61*	*NG*	*NG*		**1*	**5*		*NG*	*NG*	
SIM285	**63/*63*	**63*	**63*		**2*	**2*		**5*	**83+*2*	
SIM286	**57+*10D/*57+*10D*	**57+*10*	**57+*10*		**10*	**10*		*NG*	*NG*	
SIM287	**76+*1.002/*76+1.002*	**1*	**76+*1*		**1*	**27*		*NG*	*NG*	
SIM288	**77+*2.002/*77+*2.002*	**79+*2*	**79+*2*		**2*	**2*		*NG*	*NG*	
SIM289	**78+*2.001/*78+*2.002*	**2+*67*	**78+*2*		**2*	**2*	*(+)*	*NG*	*NG*	
SIM290	**79+*68+*4.001x2/*79+*68+*4.001x2*	**4x6*	**68-like+*80*		**4*	**4*	*(+)*	**4*	**4x5*	
SIM291	**80/*80*	*NG*	*NG*		**2*	**2*		**13C*	**13B*	
SIM292	**82.001/*82.001*	**82*	**82*		**82*	**82*		**1*	**1*	
SIM293	**10.001/*5*	**10*	**5*		**10*	**10*		**5*	**10*	
SIM294	**17.001/*5*	**17*	**5*		**17*	**17*		**5*	**17*	
SIM295	**101.001/*5*	**101*	**5*		**101*	**101*		**5*	**101*	
SIM296	**65.001/*5*	**5*	**65*		**65*	**65*		**5*	**10*	
SIM297	**24.001/*5*	**24*	**5*		**24*	**24*		**1*	**5*	
SIM298	**3.001/*5*	**3*	**5*		**3*	**3*		**3*	**5*	
SIM299	**84.001/*5*	**5*	**84*		**84*	**84*		**5*	**84*	
SIM300	**86.001/*5*	**5*	**86*		**86*	**86*		**5*	**86*	
SIM301	**10.002x3/*5*	**10*	**10+*10*		**10*	**10*		**10*	**10x2*	
SIM302	**79+*68+*4.009/*5*	**4*	**4-like*		**4*	**4*		**4*	**4*	
SIM303	**41x3/*70*	**41+*41+*41*	**70*		**41*	**70*	*(+)*	**29x2*	**39x2*	
SIM304	**2.001x2/*24.001*	**2+*2*	**24*		**2*	**24*		**1*	**2x2*	
SIM305	**35.001x2/*19.001*	**19*	**35+*35*		**19*	**35*		**19*	**35x2*	
SIM306	**43.001x2/*55.001*	**43+*43*	**55*		**43*	**55*		**43x2*	**55*	
SIM307	**29.001x2/*89.001*	**29+*29*	**89*		**29*	**89*		**1*	**29x2*	
SIM308	**17.001x2/*40.001*	**17+*17*	**40*		**17*	**40*		**17x2*	**40*	
SIM309	**10.002x3/*104.001*	**13*	**36+*10*		**10*	**104*	*(+)*	**10x2*	**39x2*	
SIM310	**10.001x2/*8.001*	**10+*10*	**8*		**8*	**10*		**8*	**10x2*	
SIM311	**79+*68+*4.009/*79+*68+*4.009*	**4+*4+*4*	**4-like*		**4*	**4*		**4*	**4x3*	

NG, not genotyped by tool due to fail error; (+), possible duplication (Astrolabe); , indicates that the call was obtained by running algorithm without BAM quality control due to previous fail; **S1+*1*, Stargazer reports this allele in cases where it does not detect all variants that it uses to define the *CYP2D6*68+*4* tandem rearrangement

Table S3.5: Astrolabe, Aldy, and Stargazer calls for samples with simulated *CYP2D6* structural variants and 100x read depth

Read depth: 100x										
Sample ID	**Truth**	**Aldy**			**Astrolabe**			**Stargazer**		
		Hap1	**Hap2**	**Notes**	**Hap1**	**Hap2**	**Notes**	**Hap1**	**Hap2**	**Notes**
SIM101	**79+*68A+*4.009/*3.001*	**3*	**4+*4-like*		**3*	**4*	*(+)*	**3*	**10x2*	
SIM102	**1.002X2+*83.001/*56.002*	**13+*1*	**61*		**1*	**56*	*(+)*	**1x3*	**56*	
SIM103	**1.001+*90/*27.001*	**1+*90*	**27*		**27*	**90*		**1*	**1x2*	
SIM104	**10.001X2/*4.001*	**10+*10*	**4*		**10*	**4*		**4*	**10x2*	
SIM105	**10.002X3/*42.001*	**10+*10+*10*	**42*		**10*	**42*	*(+)*	**10x3*	**42*	
SIM106	**76+*1.002/*5*	**1*	**76*		**1*	**27*		*NG*	*NG*	
SIM107	**79+*68A+*4.001x2/*7.001*	**4+*4+*4*	**7-like*		**7*	**4*		**4x2*	**7x2*	
SIM108	**17X2(REPdup)/*43.001 x2*	**17+*17*	**43+*43*		**17*	**43*	*(+)*	**17x2*	**43x2*	
SIM109	**17X2(spacer)/*69.001*	**17+*83*	**69*		**17*	**69*		**10*	**83+*2*	
SIM110	**1.002X3/*59.001*	**1+*1+*1*	**59*		**1*	**59*	*(+)*	**1x3*	**59*	
SIM111	**1.004X5/*4.001*	**1+*1+*1+*1+*1*	**1-like*		**1*	**1*	*(+)*	**1*	**1x5*	
SIM112	**80/*5*	**5*	**67*		*NG*	*NG*		**5*	**13B*	
SIM113	**29X2/*84.001*	**29+*29*	**84*		**29*	**84*		**29x2*	**84*	
SIM114	**2.007X2/*70.001*	**2+*2*	**70*		**2*	**70*	*(+)*	**2x2*	**39*	
SIM115	**2.010X4/*58.001*	**2+*2+*2+*2*	**58*		**2*	**58*		**2x2*	**17x3*	
SIM116	**2.001x8/*29.001x2*	**2x6*	**29+*63+*80*		**2*	**29*	*(+)*	**2x8*	**29x2*	
SIM117	**35.001X2/*95.001*	**35+*35*	**95*		**35*	**95*	*(+)*	**10*	**35x2*	
SIM118	**36.001(REP6)/*4.013*	**36*	**4*		**10*	**4*		**4*	**10*	
SIM119	**36.001+*10.001/*36.001+*10.001*	**36+*10*	**36+*10*		**10*	**10*		**36+*10*	**36+*10*	
SIM120	**36.001X2/*10.002*	**10*	**36+*10*		**10*	**10*		**10*	**36+*10*	
SIM121	**36.001X2+*10.003/*107.001*	**107+*36*	**36+*10*		**10*	**107*	*(+)*	**36+*10*	**36+*10*	
SIM122	**36.002/*3.001X2*	**3+*3*	**36*		**3*	**10*		**3x2*	**10*	
SIM123	**36+*10.001x2/*64.001*	**10+*64*	**36+*10*		**10*	**64*		*NG*	*NG*	
SIM124	**83.001/*54.001*	**39*	**54*		**1*	**54*		**10*	**39*	
SIM125	**4.013+*4.001/*12.002*	**12+*4*	**45*		**12*	**4*		**1*	**4N+*4*	
SIM126	**4.013X2/*10.004*	**36*	**4+*4*		**10*	**4*		**4N+*4*	**10*	
SIM127	**41.001X2/*68+*4.001*	**63-like*	**69*		**41*	**4*	*(+)*	**S1+*1*	**1x2*	
SIM128	**68/*4.001*	**4*	**68-like*		**4*	**4*		**S1+*1*	**13C*	
SIM129	**43.001X2/*60.001*	**43+*43*	**60*		**43*	**60*	*(+)*	**1x2*	**60*	
SIM130	**45X2/*62.001*	**45+*45*	**62*		**62*	**45*		**1x2*	**62*	
SIM131	**4.011X2/*16*	**4+*4*	**77*		**4*	**4*		**4*	**4*	
SIM132	**4.001X3/*83.003*	**36+*80*	**4*		**1*	**4*		**1*	**4x3*	
SIM133	**5/*5*	**5*	**5*		*{*2M}*	*{*2M}*		**5*	**5*	
SIM134	**57+*10D/*45.001*	**45*	**57+*10*		**10*	**45*		**10*	**83+*2*	
SIM135	**61/*4.008*	**4*	**61*		**39*	**4*		**4N+*4*	**5*	
SIM136	**63/*15.001*	**15*	**63*		**2*	**15*		*NG*	*NG*	
SIM137	**13A1/*11.001*	**1*	**79*		**1*	**11*		**11*	**13B*	
SIM138	**67/*9.001X2*	**1*	**67*		**1*	**9*		*NG*	*NG*	
SIM139	**68+*4.001/*18.001*	**1*	**4+*68-like*		**1*	**4*	*(+*68)*	**S1+*1*	**4*	
SIM140	**6.003X2/*71.001*	**6+*6*	**71*		**6*	**71*	*(+)*	**6x2*	**71*	
SIM141	**13A1+*1.004x2/*43.001*	**1+*1*	**43+*79*		**1*	**43*		*NG*	*NG*	
SIM142	**57+*10D/*35.002*	**35*	**57+*10*		**10*	**35*		**10*	**83+*2*	

Table S3.5 – continued

Sample ID	Truth	Aldy			Astrolabe			Stargazer		
		Hap1	Hap2	Notes	Hap1	Hap2	Notes	Hap1	Hap2	Notes
SIM143	*77+*2.002/*2.011	*2	*79+*2		*2	*2		NG	NG	
SIM144	*78+*2.001/*34.001	*2+*67	*34		*2	*34	(+)	*2	*S2+*1	
SIM145	*82.001/*33.001	*1	*33		*33	*82		*1	*33	
SIM146	*79+*2.001/*39.001	*39	*79+*2		*2	*4		NG	NG	
SIM147	*83.002/*5	*1	*5		*1	*5		*1	*5	
SIM148	*80/*4.004	*4	*67		*4	*4		*4	*13C	
SIM149	*29x2/*5	*29	*29		*29	*29		*29	*29	
SIM150	*1.002x4/*2.001x8	*1+*1+*61	*2x8+*79		*1	*2	(+)	*1x6	*2x6	
SIM167	*4/*68+*4	*4+*4	*68-like		*4	*4	(+*68)	*S1+*1	*4	
SIM168	*68+*4/*68+*4	*4+*68-like	*68+*4		*4	*4		*S1+*1	*S1+*1	
SIM169	*79+*68+*4.009/*2.011	*2	*4+*4		*2	*4		*4x2	*10	
SIM170	*1.002x2+*83.001/*6.001	*1+*1	*6+*61		*1	*6		*1x3	*6	
SIM171	*1.001+*90.001/*4.001	*1+*90	*4		*90	*4		*1x2	*4	
SIM172	*10.001x2/*3.001	*10+*10	*3		*3	*10		*3	*10x2	
SIM173	*3.001x2/*4.001x2	*3+*3	*4+*4		*3	*4	(+)	*3x2	*4x2	
SIM174	*76+*1.002/*34.001	*34	*76+*1		*1	*27		*S2+*1	*34	
SIM175	*79+*68+*4.001x2/*40.001	*2+*40	*4+*4		*4	*40		*2x2	*4x2	
SIM176	*17.001x2/*29.001x2	*17+*17	*29+*29		*17	*29	(+)	*17x2	*29x2	
SIM177	*17x2(spacer)/*45.001	*17+*61	*45		*17	*45		*2	*83+*2	
SIM178	*1.002x3/*2.001x2	*1+*1+*1	*2+*2		*1	*2	(+)	*1x3	*2x2	
SIM179	*80/*33.001	*33	*78		*33	*53		*13B	*33	
SIM180	*29.001x2/*4.004	*29+*29	*4		*29	*4	(+)	*4	*29x2	
SIM181	*35.001x2/*1.002	*1	*35+*35		*1	*35	(+)	*1	*2x2	
SIM182	*36.001/*9.001	*36	*9		*9	*10		*9	*10	
SIM183	*36.001x2/*100.001	*100	*36+*10		*100	*10		*36+*10	*100	
SIM184	*4.013+*4.001/*50.001	*4+*4	*50		*50	*4		*4N+*4	*50	
SIM185	*4.013x2/*51.001	*4+*4	*51		*51	*4		*4N+*4	*51	
SIM186	*43.001x2/*28.001	*28	*43+*43		*28	*43	(+)	*1x2	*28	
SIM187	*5/*5	*5	*5		{*2M}	{*2M}		*5	*5	
SIM188	*27.001/*5	*27	*5		*5	*27		*1	*5	
SIM189	*57.001+*10.003/*1.001	*1	*57+*10		*1	*10		NG	NG	
SIM190	*61/*84.001	*61	*84		*1	*84		*5	*83+*2	
SIM191	*63/*58.001	*58	*63		*2	*58		*5	*83+*2	
SIM192	*61/*1.001x3	*1+*1+*1	*61		*1	*1	(+)	NG	NG	
SIM193	*63/*2.001x2	*2+*2	*63		*2	*2		*2	*83+*2	
SIM194	*13A1/*9x2	*1	*79		*1	*9		*9	*13B	
SIM195	*67/*81	*1	*67		*1	*81		*13C	*81	
SIM196	13A+*1.004x2/*7	*1+*1	*7+*79		*1	*7		NG	NG	
SIM197	*77+*2.002/*3.001	*3	*79+*2		*2	*3		NG	NG	
SIM198	*78+*2.001/*39.001	*2+*67	*39		*2	*39		*2	*S2+*1	
SIM199	*83.002/*2.001	*2	*61		*1	*2		*2D	*39	
SIM200	*82.001/*2.011	*1	*2		*2	*82		*1	*2	
SIM201	*2.001x8/*1.001	*1	*2x7		*2	*34	(+)	*2x8	*34	
SIM202	*1.001x4/*2.001	*1+*1+*1+*1	*2		*1	*2	(+)	*1x3	*2x2	
SIM203	*2.011/*2.011	*2	*2		*2	*2		*2	*2	
SIM277	*13A+*1.004x2/*13A+*1.004x2	*1+*1+*1	*13+*1+*79		*1	*1	(+)	NG	NG	
SIM278	*13A/*13A	*13	*79		*1	*1		*13B	*13B	
SIM279	*1.001+*90.001/*1.001+*90.001	*1+*1	*90+*90		*1	*90	(+)	*1	*1x3	

Table S3.5 – continued

Sample ID	Truth	Aldy			Astrolabe			Stargazer		
		Hap1	Hap2	Notes	Hap1	Hap2	Notes	Hap1	Hap2	Notes
SIM280	**1.002X2+*83.001/*1.002x2+*83.001*	**1+*1*	**13+*1+*61*		**1*	**1*	*(+)*	**1*	**1x5*	
SIM281	**36.001/*36.001*	**10*	**36*		**10*	**10*		**10*	**10*	
SIM282	**36.001x2/*36.001x2*	**36+*10*	**36+*10*		**10*	**10*		**36+*10*	**36+*10*	
SIM283	**36.001x2+*10.003/*36x2+*10.003*	**36+*10*	**10+*10+*36+*10*		**10*	**10*	*(+)*	*NG*	*NG*	
SIM284	**61/*61*	*NG*	*NG*		**1*	**5*		*NG*	*NG*	
SIM285	**63/*63*	**63*	**63*		*{*2M}*	*{*2M}*	*(*5)*	**5*	**83+*2*	
SIM286	**57+*10D/*57+*10D*	**57+*10*	**57+*10*		**10*	**10*	*(+)*	*NG*	*NG*	
SIM287	**76+*1.002/*76+1.002*	**1+*1*	**77*		**1*	**27*		*NG*	*NG*	
SIM288	**77+*2.002/*77+*2.002*	**13+*2*	**79+*2*		**2*	**2*	*(+)*	*NG*	*NG*	
SIM289	**78+*2.001/*78+*2.002*	**78+*2*	**78+*2*		**2*	**2*	*(+)*	*NG*	*NG*	
SIM290	**79+*68+*4.001x2/*79+*68+*4.001x2*	**4+*4+*4+*4*	**4-like+*80*		**4*	**4*		**4*	**4x5*	
SIM291	**80/*80*	*NG*	*NG*		**53*	**53*		**13C*	**13B*	
SIM292	**82.001/*82.001*	**82*	**82*		**82*	**82*		**1*	**1*	
SIM293	**10.001/*5*	**10*	**5*		**5*	**10*		**5*	**10*	
SIM294	**17.001/*5*	**17*	**5*		**5*	**17*		**5*	**17*	
SIM295	**101.001/*5*	**101*	**5*		**5*	**101*		**5*	**101*	
SIM296	**65.001/*5*	**5*	**65*		**5*	**65*		**5*	**10*	
SIM297	**24.001/*5*	**24*	**5*		**5*	**24*		**1*	**5*	
SIM298	**3.001/*5*	**3*	**5*		**3*	**5*		**3*	**5*	
SIM299	**84.001/*5*	**5*	**84*		**5*	**84*		**5*	**84*	
SIM300	**86.001/*5*	**5*	**86*		**5*	**86*		**5*	**86*	
SIM301	**10.002x3/*5*	**10*	**10+*10*		**10*	**10*		*NG*	*NG*	
SIM302	**79+*68+*4.009/*5*	**4*	**4-like*		**4*	**4*		**4*	**4*	
SIM303	**41x3/*70*	**41+*41+*41*	**70*		**41*	**70*		**29x2*	**39x2*	
SIM304	**2.001x2/*24.001*	**2+*2*	**24*		**2*	**24*		**1*	**2x2*	
SIM305	**35.001x2/*19.001*	**19*	**35+*35*		**19*	**35*		**19*	**35x2*	
SIM306	**43.001x2/*55.001*	**43+*43*	**55*		**43*	**55*		**43x2*	**55*	
SIM307	**29.001x2/*89.001*	**29+*29*	**89*		**29*	**89*		**1*	**29x2*	
SIM308	**17.001x2/*40.001*	**17+*17*	**40*		**17*	**40*		**17x2*	**40*	
SIM309	**10.002x3/*104.001*	**10+*10+*10*	**104*		**10*	**104*		**2*	**10x3*	
SIM310	**10.001x2/*8.001*	**10+*10*	**8*		**8*	**10*		**8*	**10x2*	
SIM311	**79+*68+*4.009/*79+*68+*4.009*	**4+*4+*4*	**4-like*		**4*	**4*		**4*	**4x3*	

NG, not genotyped by tool due to fail error; (+), possible duplication (Astrolabe); , indicate that the call was obtained by running algorithm without BAM quality control due to previous fail; **S1+*1*, Stargazer reports this allele in cases where it does not detect all variants that it uses to define the *CYP2D6*68+*4* tandem rearrangement

Table S3.6: Genotypes called by Astrolabe, Aldy, Stargazer and their ensemble in comparison to GeT-RM consensus calls

Sample	Truth			Astrolabe			Aldy			Stargazer			Ensemble			SP
	Diplotype	Activity score	Phenotype	Diplotype	Activity score	Phenotype	Diplotype	Activity score	Phenotype	Diplotype	Activity score	Phenotype	Diplotype	Activity score	Phenotype	
NA18484	*1/*17	1.5	NM	*1/*17	1.5	NM	*61-like/*67	0	PM	*1/*17	1.5	NM	*1/*17	1.5	NM	AFR
NA18509	*2/*17	1.5	NM	*2/*17	1.5	NM	*17/*2	1.5	NM	*2/*17	1.5	NM	*2/*17	1.5	NM	AFR
NA18518	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM	AFR
NA18519	*1/*29	1.5	NM	*29/*106	NA	No call (IM/NM)	*106/*29	NA	No call (IM/NM)	*29/*106	NA	No call (IM/NM)	*29/*106	NA	No call (IM/NM)	AFR
NA18855	*1/*5	1	IM	*5/*27	1	IM	*1/*5	1	IM	*1/*5	1	IM	*1/*5	1	IM	AFR
NA18861	*5/*29	0.5	IM	*5/*29	0.5	IM	*29/*5	0.5	IM	*5/*29	0.5	IM	*5/*29	0.5	IM	AFR
NA18868	*2/*5	1	IM	*2/*2	2	NM	*2/*5	1	IM	*2/*5	1	IM	*2/*5	1	IM	AFR
NA19095	*1/*29	1.5	NM	*1/*29	1.5	NM	*1/*29	1.5	NM	*1/*29	1.5	NM	*1/*29	1.5	NM	AFR
NA19109	*2x2/*29	2.5	UM	*2/*29	1.5	NM	*2x2/*29	2.5	UM	*2x2/*29	2.5	UM	*2x2/*29	2.5	UM	AFR
NA19122	*2/*17	1.5	NM	*2/*17	1.5	NM	*17/*2	1.5	NM	*2/*17	1.5	NM	*2/*17	1.5	NM	AFR
NA19143	*2(*45)/*10	1.25	NM	*10/*45	1.25	NM	*10/*45	1.25	NM	*2(*45)/*10	1.25	NM	*10/*45	1.25	NM	AFR
NA19147	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM	*17/*39	1.5	NM	*17/*29	1	IM	AFR
NA19174	*4/*40	0	PM	*4/*40	0	PM	*4/*40	0	PM	*4/*40	0	PM	*4/*40	0	PM	AFR
NA19176	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	AFR
NA19178	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	AFR
NA19207	*2x2/*10	2.25	NM	*2/*10	1.25	NM	*2x2/*10	2.25	NM	*10/*2x2	2.25	NM	*2x2/*10	2.25	NM	AFR
NA19213	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	AFR
NA19226	*2/*2x2	3	UM	*2/*2 (+)	3	UM	*2/*2x2	3	UM	*2/*2x2	3	UM	*2x2/*2	3	UM	AFR
NA19238	*1/*17	1.5	NM	*1/*17	1.5	NM	*1/*17	1.5	NM	*1/*17	1.5	NM	*1/*17	1.5	NM	AFR
NA19239	*15/*17	0.5	IM	*15/*17	0.5	IM	*15/*17	0.5	IM	*1/*15	1.0	IM	*15/*17	0.5	IM	AFR
NA19240	*15/*17	0.5	IM	*15/*17	0.5	IM	*15/*17	0.5	IM	*1/*15	1.0	IM	*15/*17	0.5	IM	AFR
HG01190	*5/*68+*4	0	PM	*4 /*4 (+*68), (*5)	0	PM	*4/*68	0	PM	*5/*68+*4	0	PM	*5/*68+*4	0	PM	AMR
NA19789	*1/*1	2	NM	*1/*1	2	NM	*1/*61	1	IM	*1/*1	2	NM	*1/*1	2	NM	AMR
NA19819	*2/*4x2	1	IM	*2/*4	1	IM	*2/*4x2	1	IM	*2/*4x2	1	IM	*2/*4x2	1	IM	AMR
NA19908	*1/*46	2	NM	*43/*45	NA	No call (IM/NM)	*1/*46	2	NM	*1/*2	2	NM	[mac]	NA	No call	AMR
NA19917	*1/*40	1	IM	*1/*40	1	IM	*1/*40	1	IM	*1/*40	1	IM	*1/*40	1	IM	AMR
NA19920	*1/*4x2	1	IM	*1/*4	1	IM	*1/*4x2	1	IM	*1/*4x2	1	IM	*1/*4x2	1	IM	AMR
NA20296	*1/*2	2	NM	*1/*2	2	NM	*63/*67	NA	No call	*1/*2D	2	NM	*1/*2	2	NM	AMR
HG00436	*2x2/*71	NA	No call (NM/UM)	*2/*71	NA	No call (IM/NM)	*2x2/*71	NA	No call (NM/UM)	*71/*2x2	NA	No call (NM/UM)	*2x2/*71	NA	No call	EAS
HG00589	*1/*21	1	IM	*1/*2	2	NM	*1/*21	1	IM	*1/*21	1	IM	*1/*2	1	IM	EAS
NA18524	*1/*36x2+*10	1.25	NM	*1/*10	1.25	NM	*1+*36/*36+*10	1.25	NM	*36+*10/*36+*10	0.5	IM	*1/*10(+*36x2)	1.25	NM	EAS
NA18526	*1/*36x2+*10	1.25	NM	*1/*10	1.25	NM	*1+*36/*36+*10	1.25	NM	*36+*10/*36+*10	0.5	IM	*1/*10(+*36x2)	1.25	NM	EAS
NA18540	(*36+)*10/*41	0.75	IM	*41/*10 (+)	0.75	IM	*36+*10/*61+*69	0.25	IM	*36+*10/*36+*10	0.5	IM	[mac]	NA	No call	EAS
NA18544	*10/*41	0.75	IM	*10/*41	0.75	IM	*10/*41	0.75	IM	*10/*41	0.75	IM	*10/*41	0.75	IM	EAS
NA18552	*1/*14	1.5	NM	*1/*2	2	NM	*1/*14	1.5	NM	*1/*14	1.5	NM	*1/*14	1.5	NM	EAS
NA18564	*2/*36+*10	1.25	NM	*2/*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	EAS
NA18565	*10/*36x2	0.25	IM	*10/*10	0.5	IM	*10/*36+*10	0.5	IM	*10/*36+*10	0.5	IM	*10/*36+*10	0.25	IM	EAS
NA18617	*36+*10/*36+*10	0.5	IM	*10/*10 (+)	0.75	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	EAS
NA18942	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	EAS
NA18952	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	EAS
NA18959	*2/*36+*10	1.25	NM	*2/*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	EAS
NA18966	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	EAS
NA18973	*1/*21	1	IM	*1/*2	2	NM	*1/*21	1	IM	*1/*21	1	IM	*1/*21	1	IM	EAS
NA18980	*2/*36+*10	1.25	NM	*2/*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	EAS
NA18992	*1/*5	1	IM	*1/*5	1	IM	*1/*5	1	IM	*1/*5	1	IM	*1/*5	1	IM	EAS
NA19003	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	EAS
NA19007	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	EAS
HG00276	*4/*5	0	PM	*4/*4	0	PM	*4/*5	0	PM	*4/*5	0	PM	*4/*5	0	PM	EUR
NA06991	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	EUR
NA07000	*9/*2(*35)	1.5	NM	*9/*35	1.5	NM	*35/*9	1.5	NM	*9/*35	1.5	NM	*9/*35	1.5	NM	EUR
NA07019	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	EUR
NA07029	*1/*35	2	NM	*1/*35	2	NM	*1/*35	2	NM	*1/*35	2	NM	*1/*35	2	NM	EUR
NA07055	*4/*4	0	PM	*4/*4	0	PM	*4/*4	0	PM	*4/*4	0	PM	*4/*4	0	PM	EUR
NA07056	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	EUR
NA07348	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM	EUR
NA07357	*1/*6	1	IM	*5/*6	0	PM	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM	EUR
NA10831	*5/*4	0	PM	*5/*4	0	PM	*4/*5	0	PM	*5/*4	0	PM	*5/*4	0	PM	EUR

Table S3.6 – continued

Sample	Truth			Astrolabe			Aldy			Stargazer			Ensemble			SP
	Diplotype	Activity score	Phenotype	Diplotype	Activity score	Phenotype	Diplotype	Activity score	Phenotype	Diplotype	Activity score	Phenotype	Diplotype	Activity score	Phenotype	
NA10847	*1/*41	1.5	NM	*1/*41	1.5	NM	*1/*41	1.5	NM	*1/*41	1.5	NM	*1/*41	1.5	NM	EUR
NA10851	*1/*4	1.5	NM	*1/*4	1.5	NM	*1/*4	1.5	NM	*1/*4	1.5	NM	*1/*4	1.5	NM	EUR
NA10854	*1/*4	1.5	NM	*1/*4	1.5	NM	*1/*4	1.5	NM	*1/*4	1.5	NM	*1/*4	1.5	NM	EUR
NA11832	*1/(*68)+*4	1	IM	*1/*4 (+*68)	1	IM	*1/*68+*4	1	IM	*1/*68+*4	1	IM	*1/*68+*4	1	IM	EUR
NA11839	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	EUR
NA11993	*1/*9	1.5	NM	*1/*9	1.5	NM	*1/*9	1.5	NM	*1/*9	1.5	NM	*1/*9	1.5	NM	EUR
NA12003	*4/*35	1	IM	*4/*35 (+)	1	IM	*35/*4	1	IM	*4/*35	1	IM	*35/*4	1	IM	EUR
NA12006	*4/*41	0.5	IM	*4/*41	0.5	IM	*4/*41	0.5	IM	*4/*41	0.5	IM	*4/*41	0.5	IM	EUR
NA12145	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	EUR
NA12156	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	EUR
NA12717	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	EUR
NA12813	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	EUR
NA12873	*1/*5	1	IM	*1/*5	1	IM	NG	NA	No call	*1/*5	1	IM	*1/*5	1	IM	EUR
NA20509	*4/*35	2	NM	*4/*35	2	NM	*35/*4	2	NM	*4/*35	2	NM	*4/*35	2	NM	EUR
NA21781	*2x2/*68+*4	2	NM	*2/*4 (+)	1 or 2	No call (IM/NM)	*2x2/*68+*4	2	NM	*2x2/*68+*4	2	NM	*2x2/*68+*4	2	NM	EUR
NA12892	*2/*3	1	IM	*2/*3	1	IM	*2/*3	1	IM	*2/*3	1	IM	*2/*3	1	IM	EUR
NA12891	*41/*68+*4	0.5	IM	*41/*4(+*68)	0.5	IM	*41/*68+*4	0.5	IM	*41/*68+*4	0.5	IM	*41/*68+*4	0.5	IM	EUR
NA12878	*3/*68+*4	0	PM	*3/*4(+*68)	0	PM	*3/*68+*4	0	PM	*3/*68+*4	0	PM	*3/*68+*4	0	PM	EUR

NG, not genotyped; [mac], ambiguous call; (+), possible duplication (Astrolabe); PM, poor metaboliser (AS = 0); IM, intermediate metaboliser (0 < AS < 1.25); NM, normal metaboliser (1.25 $\leq$ AS $\leq$ 2.25); UM, ultrarapid metaboliser (AS > 2.25)
Calls written in black text are concordant to the truth set; red colour indicates discordant haplotypes/diplotypes; blue colour indicates discordant activity scores and discordant phenotypes; teal colour indicates correct gene copies but discordant arrangement reported (Aldy)
NA18519 — GeT-RM genotyped this sample as *1/*29. However, due to the presence of 3878G>A from NGS, *106/*29 is considered to be the true positive call for this analysis.

Table S3.7: Genes that can be genotyped using Astrolabe, Aldy, and Stargazer

Astrolabe (v0.8.7.0)	Aldy (v2.2.3)	Stargazer (v1.0.7)
CYP2D6	*CYP2D6*	*CYP2D6*
CYP2C9	*CYP2C9*	*CYP2C9*
CYP2C19	*CYP2C19*	*CYP2C19*
	CYP2C8	*CYP2C8*
	CYP3A4	*CYP3A4*
	CYP3A5	*CYP3A5*
	CYP4F2	*CYP4F2*
	CYP2A6	*CYP2A6*
	TPMT	*CYP2B6*
	DPYD	*CYP1A1*
		CYP1A2
		CYP1B1
		CYP2A13
		CYP2E1
		CYP2F1
		CYP2J2
		CYP2R1
		CYP2S1
		CYP2W1
		CYP3A7
		CYP3A43
		CYP4B1
		CYP26A1
		CYP19A1
		TPMT
		DPYD
		G6PD
		GSTM1
		GSTP1
		GSTT1
		NAT1
		NAT2
		NUDT15
		POR
		SLC15A2
		SLC22A2
		SLC01B1
		SLC02B1
		TBXAS1
		UGT1A1
		UGT2B7
		UGT2B15
		UGT2B17
		VKORC1

This information is adapted from the following repositories (accessed 20-2-2020);
`https://www.childrensmercy.org/genomesoftwareportal/`
`https://github.com/inumanag/aldy`
`https://stargazer.gs.washington.edu/stargazerweb/`

Chapter 4

StellarPGx: A Nextflow pipeline for calling star alleles in cytochrome P450 genes

The work presented in this chapter was published in *Clinical Pharmacology and Therapeutics* (see Appendix A and Appendix E), and has been reprinted with permission from Wiley.

Twesigomwe D., Drögemöller B.I., Wright G.E.B., Siddiqui A., da Rocha J., Lombard Z. and Hazelhurst S. (2021). StellarPGx: A Nextflow pipeline for calling star alleles in cytochrome P450 genes. *Clinical Pharmacology and Therapeutics* 110(3), 741-749. `https://doi.org/10.1002/cpt.2173`.

4.1 Summary

Bioinformatics pipelines for calling star alleles (haplotypes) in cytochrome P450 (CYP) genes are important for the implementation of precision medicine. Genotyping CYP genes using high throughput sequencing data is complicated, e.g. by being highly polymorphic, not to mention the structural variations especially in *CYP2D6*, *CYP2A6*, and *CYP2B6*. Genome graph-based variant detection approaches have been shown to be reliable for genotyping *HLA* alleles. However, their application to enhancing star allele calling in CYP genes has not been extensively explored. We present StellarPGx, a Nextflow pipeline for accurately genotyping CYP genes by combining genome graph-based variant detection, read coverage information from the original reference-based alignments, and combinatorial diplotype assignments. The implementation of StellarPGx using Nextflow facilitates its portability, reproducibility, and scalability on various user platforms. StellarPGx is currently able to genotype 12 important pharmacogenes belonging to the *CYP1*, *2*, and *3* families. For purposes of validation, we use *CYP2D6* as a model gene owing to its high degree of polymorphisms (over 130 star alleles defined to date, including complex structural variants) and clinical importance. We applied StellarPGx and three existing callers to 109 whole genome sequenced samples for which the Genetic Testing Reference Material Coordination Program (GeT-RM) has recently provided consensus truth *CYP2D6* diplotypes. StellarPGx had the highest *CYP2D6* diplotype concordance (99%) with GeT-RM compared with Cyrius (98%), Aldy (82%), and Stargazer (84%). This exemplifies the high accuracy of StellarPGx and highlights its importance for both research and clinical pharmacogenomics applications. The StellarPGx pipeline is open-source and available from `https://github.com/SBIMB/StellarPGx`.

4.2 Introduction

The human cytochrome P450 (CYP) gene family currently comprises 57 genes. Of these, 12 (belonging to the *CYP1*, *2*, and *3* families) encode enzymes responsible for the bioactivation or elimination of over 80% of commonly prescribed medications. These CYP genes are also known to be highly polymorphic (Zanger and Schwab, 2013). Accordingly, they account for the highest number of pharmacogenetic biomarkers in drug labels according to the FDA (`https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm`). Genotyping CYP genes and subsequent phenotype prediction is therefore a fundamental part of precision medicine approaches as CYP allele information is used for adjusting therapy, thus countering drug inefficacy and adverse drug reactions (Goetz et al., 2018; Rajman et al., 2017; Y. Zhou et al., 2017).

CYP gene haplotypes — single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and/or structural variations that are inherited together — are commonly denoted using the star (*) allele nomenclature to aid in interpretation (Caudle et al., 2020). We refer to SNPs and indels collectively as small nucleotide variations (SNVs) for the rest of this manuscript. The metaboliser activity of star alleles for most CYP genes ranges from non-functional to increased function, making it possible to predict individual phenotypes from diplotypes via activity score systems (Caudle et al., 2020; Gaedigk, Dinh, et al., 2018; Gaedigk et al., 2008; Zubiaur et al., 2020). This makes accurate genotyping of CYP genes very useful and imperative. Notably, the Pharmacogene Variation Consortium (PharmVar) (Gaedigk, Ingelman-Sundberg, et al., 2018; Gaedigk, Sangkuhl, et al., 2019) and the Pharmacogenomics Knowledgebase (PharmGKB) (Whirl-Carrillo et al., 2012) provide resourceful, curated catalogues of the haplotype information for CYP genes (and other pharmacogenes) as well as links to the related clinical annotations, respectively.

The ever-decreasing cost and increasing ubiquity of high-throughput sequencing (HTS) provide the opportunity for clinical-grade whole genome sequencing (WGS) and extensive characterisation of CYP allele distribution especially in previously understudied populations. However, single molecule long-read technologies, which would make haplotype detection straight-forward (Ammar et al., 2015; Buermans et al., 2017; Liau et al., 2019; Qiao et al., 2016), are more expensive and less accessible compared to short-read technologies such as Illumina (D. R. Bentley et al., 2008). Notably, calling star alleles in CYP genes using short-read Next Generation Sequencing (NGS) data is challenging especially for genes which have neighbouring pseudogenes (e.g. *CYP2D6*, *CYP2B6* and *CYP2A6*) or isoforms (e.g. *CYP2C19* and *CYP2C9*) with high sequence similarity, hence causing a number of structural rearrangements, sequencing problems and subsequent read misalignments (Nofziger and Paulmichl, 2018; Yang et al., 2017).

The current state-of-the-art bioinformatics algorithms for calling star alleles in CYP and/or other

pharmacogenes from short-read NGS data include Astrolabe (formerly Constellation) (Twist et al., 2016), Aldy (Numanagić et al., 2018), Stargazer (Lee et al., 2019a; Lee et al., 2019b), PharmCAT (Sangkuhl et al., 2020), and Cyrius (Chen et al., 2021). These algorithms mainly use Binary alignment Map (BAM) files and/or Variant Call Format (VCF) files as input to call SNV-defined alleles as well as structural variants. A notable commonality among these existing tools is that they all rely on variant detection based on reads mapped to a standard reference genome (H. Li and Durbin, 2009; Wu and Nacu, 2010) (Genome Reference Consortium Human Build 37 (GRCh37) or GRCh38) as a fundamental step in their diplotype calling. Although this approach has facilitated variant calling from HTS data over the years, it has been shown to have limitations such as reference allele bias and read misalignment around indels (Ambler et al., 2019; Garrison et al., 2018; H. Li et al., 2020). Aligning reads to a reference genome without awareness of variation could therefore lead to erroneous variant calling and subsequently inaccurate star allele calling.

There is a growing paradigm shift towards using pangenomes to counter the limitations of reference-based alignments. A pangenome is a graph structure characterised by having a linear reference genome augmented with prior variant information such that each haplotype is represented as a path in the graph (Ambler et al., 2019; Garrison et al., 2018; H. Li et al., 2020). This facilitates more accurate read alignment as well as simultaneous genotyping of variants present in the graph (Rakocevic et al., 2019). Furthermore, novel variants can still be accurately called and used to update the graph. Given the high degree of polymorphisms and complex variation in most CYP genes, they are an excellent application for pangenome-based alignment and genotyping approaches. High genotyping accuracy, courtesy of genome graphs, has already been shown for similarly complex genomic regions such as the human leukocyte antigen (*HLA*) complex (Eggertsson et al., 2017; Kim et al., 2019) and short tandem repeat regions (Dolzhenko et al., 2019).

Here we present StellarPGx, a highly efficient hybrid pipeline that combines genome graph-based variant detection, read coverage information from the original HTS data alignments, and combinatorial diplotype assignment for genotyping CYP genes in a reproducible manner. StellarPGx is implemented using Nextflow (Di Tommaso et al., 2017), an in silico workflow management system that facilitates parallelisation, scalability, reproducibility and portability of pipelines via Docker (`https://www.docker.com`) and Singularity (Kurtzer et al., 2017) technologies. Docker and Singularity allow for the easy distribution of all the software dependencies for StellarPGx in a containerised system thus overcoming reproducibility issues caused by the diversity in user computational frameworks.

We use *CYP2D6* as a model gene for validating StellarPGx using WGS data from 109 DNA samples that have been characterised as part of the Centers for Disease Control and Prevention's

(CDC) Genetic Testing Reference Materials Coordination Program (GeT-RM) (Gaedigk, Turner, et al., 2019). *CYP2D6* provides the ideal use case for StellarPGx as over 130 star alleles have been catalogued by PharmVar (Gaedigk, Ingelman-Sundberg, et al., 2018), including complex structural rearrangements with the neighbouring non-functional paralog, *CYP2D7*. GeT-RM recently recharacterised 179 samples to resolve truth diplotypes for *CYP2D6* based on consensus results from various laboratories that used orthogonal genotyping techniques such as XL-PCR, Sanger sequencing, PacBio sequencing and TaqMan CNV analysis (Gaedigk, Turner, et al., 2019). One hundred nine of the GeT-RM samples have publicly available short-read WGS data covering over 40 *CYP2D6* star alleles, which we used for the validation of StellarPGx and comparison with Aldy and Stargazer, which we found to be currently the best performing tools according to our previous benchmark (Chapter 3), as well as Cyrius (Chen et al., 2021) (developed more recently).

4.3 Methods

4.3.1 Implementation

We developed StellarPGx using Nextflow (Di Tommaso et al., 2017) a domain-specific language for creating scalable, highly parallelisable and portable pipelines with components that can be written in various programming languages. StellarPGx is containerised using Docker and Singularity for easy distribution of all the required software dependencies thus facilitating workflow reproducibility on user systems. The major steps in the StellarPGx pipeline are summarised in Figure 4.1 and described below using *CYP2D6* as a model gene. The StellarPGx version described here is currently able to identify *CYP2D6*1* through *CYP2D6*141*.

4.3.2 Input data and graph-based variant calling

StellarPGx takes BAM or CRAM files as input since they are the most common formats for storing and sharing WGS data that can be used for both SNV calling and copy number determination. StellarPGx supports alignments to both GRCh37 and GRCh38.

Typical steps in pangenome-based variant detection (via the variation graph route rather than assembly) include genome graph construction using a reference genome and input variants (usually common variants in a population), graph indexing, alignment of sequence reads to the pangenome and subsequent variant calling (Garrison et al., 2018). Rather than reinventing algorithms for these initial steps, StellarPGx uses GraphTyper (Eggertsson et al., 2019; Eggertsson et al., 2017). GraphTyper is able to retrieve sequence reads from a defined region in a BAM/CRAM file, and realign them to a variant-aware graph (pangenome) thus supporting SNV and structural variant genotyping based on the paths the reads align to and alignment coverage or breakpoint

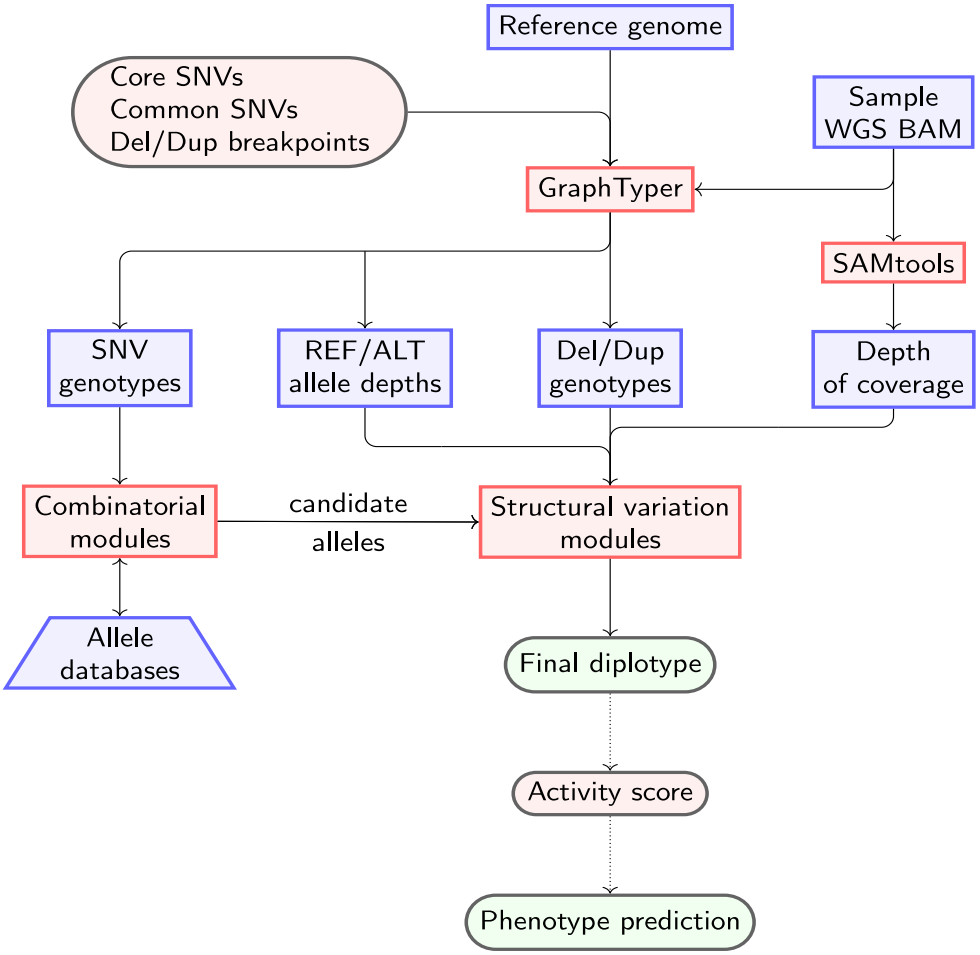


Figure 4.1: Overall summary of the StellarPGx workflow. The supported input in the version described here is a WGS (whole genome sequence) BAM (binary alignment map) file. Common and allele-defining SNVs (small nucleotide variations) as well as Del (gene deletion) and Dup (duplication) breakpoints are supplied in Variant Call Format for GraphTyper's pangenome construction. Copy number sentinel regions (target gene and control loci) are supplied in a BED (browser extensible data) file for read depth computation. StellarPGx's Nextflow processes then use the output from GraphTyper and SAMtools bedcov as well as input from allele/suballele combination databases to call sample diplotypes for a target cytochrome P450 (CYP) gene via Python 3.8 custom modules. Subsequent steps include determining the AS (activity score) based on the genotype followed by phenotype prediction depending on the AS group. These last two steps have not yet been implemented for all the supported CYP genes as consensus AS and/or phenotype assignment recommendations are available for only a few genes such as *CYP2D6*, *CYP2B6*, *CYP2C19*, and *CYP2C9*. ALT, alternate allele; REF, reference allele.

genotypes, respectively. StellarPGx uses GraphTyper to realign all reads from *CYP2D6* and *CYP2D7* to a genome graph with known *CYP2D6* variants, and simultaneously return variant genotypes as well as allele depth information in VCF format for each sample (see Appendix D on page 223 for an illustration of this process). Importantly, GraphTyper output also comprises novel sample variants.

The default prior variants included in StellarPGx for constructing the pangenome include all the

allele-defining *CYP2D6* SNVs, as well as variants with a global frequency above 1% according to either the 1000 Genomes Project (Auton et al., 2015) or gnomAD (Karczewski et al., 2018) databases. For genotyping the *CYP2D6* full gene deletion and gene duplication via GraphTyper, StellarPGx uses a VCF with predefined imprecise breakpoints (see Table 4.1 on page 91).

Table 4.1: Imprecise breakpoints used by StellarPGx for GraphTyper's gene deletion/duplication genotyping.

Gene	GRCh37 breakpoint [a] / REF allele	GRCh38 breakpoint [a] / REF allele
CYP2D6	chr22:42522000 / T	chr22:42126000 / T
CYP2A6	chr19:41349441 / C	chr19:40843536 / C
CYP2B6	chr19:41497138 / G	chr19:40991282 / G
CYP2C19	chr10:96522463 / A	chr10:94762681 / A
CYP2E1	chr10:135340300 / T	chr10:133527363 / T

[a] Note that these breakpoints are based on example in-house observations from publicly available WGS short-read data i.e. they are not experimentally validated, hence the term 'imprecise breakpoints'. The genes included here are more prone to copy number variations compared to other cytochrome P450 (CYP) genes.

4.3.3 Star allele calling

StellarPGx's SNV-defined allele caller is written in Python 3.8 (`https://docs.python.org/3.8/`) It identifies known allele-defining variants from the VCF produced by GraphTyper and matches them to a plain-text StellarPGx database developed from PharmVar's star allele definitions (Gaedigk, Ingelman-Sundberg, et al., 2018). The StellarPGx database contains all possible combinations of known SNV-defined alleles for *CYP2D6* and other supported CYP genes. For most cases, a combination of core SNVs corresponds to one diplotype. In such a scenario, this diplotype is considered as the candidate diplotype in the sample. For the cases where the presence of a particular combination of core SNVs matches more than one candidate diplotype (typically 2 or 3), we query a StellarPGx sub-database with suballele combinations to find the candidate suballele-resolved combination that best explains all the SNV genotypes in the sample. This two-step process helps identify star alleles faster especially for cases where there is only one candidate diplotype. However, there are some candidate diplotypes which produce exactly the same variant combinations even when sub-variants are considered. For example, *CYP2D6*103.001/ *3.001* and *CYP2D6*102.001/*3.002* produce a VCF with exactly the same variants as well as their zygosity (see Appendix D, page 225). StellarPGx calls *CYP2D6*102/*3* or **103/*3* for a sample with such a combination. Such cases arise mostly due to alleles with a 'limited' PharmVar level of evidence (Gaedigk, Ingelman-Sundberg, et al., 2018), meaning that their definitions are based on limited information such as exon-only sequencing as opposed to full gene resequencing.

The next step in the star allele calling process is discriminating structural variant alleles, if

present. Due to the many *CYP2D6* structural variant rearrangements (Gaedigk, Ingelman-Sundberg, et al., 2018), we integrate read coverage data from the original reference-based WGS BAM/CRAM file as well to support calling of complex *CYP2D6* structural rearrangements with *CYP2D7*. We use SAMtools (H. Li et al., 2009) to compute the depth of coverage for predefined *CYP2D6*, *CYP2D7*, and control gene regions as well as exons/introns affected by known hybrid rearrangements. For example, exon 9 is a priority as there are a number of *CYP2D6-2D7* alleles with an exon 9 conversion including *CYP2D6*36*, **57*, and **83* (Gaedigk, Ingelman-Sundberg, et al., 2018). Details for all the GRCh37 and GRCh38 regions used in StellarPGx's structural variant calling can be found in the StellarPGx project repository (`https://github.com/SBIMB/StellarPGx`). Vitamin D Receptor (*VDR*) and Epidermal Growth Factor Receptor (*EGFR*) genes have been shown to be effective as control genes in star allele calling (Lee et al., 2019) and they have relatively low CNV frequencies (Collins et al., 2020; see also `https://gnomad.broadinstitute.org/gene/ENSG00000146648?dataset=gnomad_sv_r2_1`; `https://gnomad.broadinstitute.org/gene/ENSG00000111424?dataset=gnomad_sv_r2_1`) hence we use the average coverage of both genes as a control for copy number. The implementation of StellarPGx in Nextflow allows for this depth of coverage computation to be performed in parallel with the aforementioned pangenome-based variant calling.

We determine the total copy number of *CYP2D6* alleles in the sample by comparing the read coverage for *CYP2D6* and its flanking regions with the read coverage for the control loci. If a copy number of 2 is ascertained, StellarPGx reports the candidate SNV diplotypes as the final result except for cases where we check for presence of possible singleton hybrids. For example, we use individual algorithms to check whether *CYP2D6*36* (exon 9 conversion) is present instead of the initially called *CYP2D6*10*. If a copy number of 1 is detected, StellarPGx replaces the homozygous diplotype assignments from the SNV information with a *CYP2D6 *5/*<allele>* assignment (**5* represents the *CYP2D6* gene deletion allele). For example, a sample with *CYP2D6*5/*17* would initially be genotyped as *CYP2D6*17/*17* based on SNVs alone. After running the SV caller algorithm, StellarPGx would then report the final diplotype as *CYP2D6*5/*17*. The detection of copy number 0 would default to a *CYP2D6*5/*5* diplotype call.

On the other hand, if a copy number of 3 or more is detected, StellarPGx uses the allele depth ratios (from GraphTyper VCF output) for each core SNV to decipher which of the two alleles in a sample is duplicated/multiplicated. For duplications involving tandem rearrangements, StellarPGx runs hybrid caller modules to discriminate the hybrid copies based on information from PharmVar structural variant definitions (Gaedigk, Ingelman-Sundberg, et al., 2018). For example, let us consider the *CYP2D6*68+*4* tandem (found frequently in Europeans) which comprises the non-functional **4* allele and the *CYP2D6-2D7 *68* hybrid (intron 1 switch region). StellarPGx ascertains the *CYP2D6*68+*4* tandem in a sample with the *CYP2D6*41/*68+*4*

diplotype by comparing the coverage from the promoter region to intron 1 versus the coverage from intron 1 to exon 9 as well as the allele depth (and zygosity) for 100C>T (found in *CYP2D6*68* and most *CYP2D6*4* alleles) versus that for 1847G>A (**4*) (See Figure 4.2b on page 96 for an example case). Other significantly frequent tandems include *CYP2D6*36+*10*, **36x2* and **36x2+*10* which can be discriminated from *CYP2D6*10x2* and **10x3* by interrogating the *CYP2D6* exon 9 coverage (See Figure 4.2c on page 96). In some cases (such as *CYP2D6*13* hybrid alleles — see Figure 4.2d on page 96) the coverage of CYP2D7 regions is also considered to enhance StellarPGx's recall. More details about the modules for calling the wide array of CYP structural variants (especially for *CYP2D6*) can be found in the StellarPGx project repository (`https://github.com/SBIMB/StellarPGx`).

4.3.4 StellarPGx output

The StellarPGx output file contains the final diplotype call and phenotype assignment (where applicable) of the target CYP gene as well as a breakdown of the variants and copy number information used for diplotype assignment. Other candidate solutions that were considered by the caller (if applicable) are also provided as this information could be useful especially when performing ensemble star allele calling (using at least 3 tools to enhance accuracy (see Chapter 3). Example output files are provided in Appendix D on page 221.

For cases where StellarPGx detects a novel allele or suballele, it will indicate "Possible novel major allele or suballele present: interpret with caution; Experimental validation and expert review through PharmVar is recommended" in the output (see Appendix D, page 222). Novel allele discovery is more or less straight-forward for samples with novel combinations of existing core variants. However, detecting novel alleles when completely novel core variants are present in a sample is more challenging. The current strategy is to include novel (i.e. not yet in the PharmVar database) missense, stop loss, stop gain, and frameshift CYP variants in the StellarPGx database so that can be flagged by StellarPGx when they appear in the GraphTyper-generated VCF. These variants include SNVs identified from large-scale genomics projects such as the 1000 genomes project (Auton et al., 2015) and gnomAD (Karczewski et al., 2018), as well as possible missense variants that could occur in the target CYP genes even though they haven't been identified in previous studies. However, we are yet to include support for calling novel structural variants. The structural variants that are currently supported by StellarPGx are those defined by PharmVar as per the v4.2.4 update (Gaedigk, Ingelman-Sundberg, et al., 2018). We strongly recommend that users liaise with PharmVar for information about experimental validation of possible novel star alleles and/or their designation if any of these cases arise in clinical or research settings.

4.3.5 Validation

In order to validate the StellarPGx pipeline, we performed *CYP2D6* star allele calling on publicly available high coverage WGS data for 109 ethnically diverse reference samples from the GeT-RM project. This includes 75 samples from the previous GeT-RM work (Pratt et al., 2016) and 34 of the 42 new samples added in the latest GeT-RM study (Gaedigk, Turner, et al., 2019), which was focused on recharacterisation of CYP2D6 alleles for 179 DNA samples (not all have WGS data publicly available) derived from Coriell cell lines. (see Table 4.2 for details about the data sources).

Table 4.2: Whole genome sequence data sources

Datasets	**Source**
NA19790, NA10860 and HG00423	`https://www.ncbi.nlm.nih.gov/bioproject/PRJEB36890`
31 samples (from 1000 Genomes Project) corresponding to GeT-RM reference materials	`https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31736`
70 GeT-RM samples (Polaris HiSeqX Pharmacogenomics Cohort)	Data: `https://www.ncbi.nlm.nih.gov/bioproject/PRJEB19931` Metadata: `https://github.com/Illumina/Polaris`
NA12878, NA12891, NA12892 and Yoruba Y117 trio (NA19238, NA19239, NA19240)	`http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/`

We compared CYP2D6 star allele calls and phenotype assignments by StellarPGx (version described here covers *CYP2D6*1–*141*) to the GeT-RM consensus reference materials (cover *CYP2D6*1–*114*), Aldy v2.2.6 (covers *CYP2D6*1–*114*), Stargazer v1.0.8 (covers *CYP2D6*1–*138*), and Cyrius v1.1 (covers CYP2D6*1 – *139). For Aldy and Stargazer, we performed read realignment (CYP2D and control regions) for 34 of the samples from GRCh38 to GRCh37 using BWA-MEM (H. Li and Durbin, 2009). We used the *VDR* gene as the control region for the Stargazer run. The control regions for Aldy and Cyrius are predetermined within their algorithms.

Furthermore, we genotyped 11 other CYP genes (including *CYP2A6*, *CYP2B6*, *CYP2C19*, *CYP2C9*, *CYP2C8*, *CYP3A4*, *CYP3A5*, *CYP2A1*, *CYP1A2*, *CYP2E1* and *CYP4F2*) using 75 WGS samples corresponding to reference materials from the GeT-RM study by Pratt et al. (2016), to demonstrate the utility of StellarPGx in characterising the diverse CYP allele landscape. However, for some genes, not all the samples had corresponding GeT-RM consensus calls reported. Therefore, StellarPGx's concordance was determined using the samples for which the GeT-RM consensus calls were available.

Table 4.3: Summary of the *CYP2D6* diplotype concordance for StellarPGx in comparison to other tools with respect to the 109 GeT-RM WGS samples used in the study.

Allele caller	Concordant samples	Concordance
StellarPGx	108	99%
Cyrius	107	98%
Aldy	89	82%
Stargazer	92	84%

4.4 Results

4.4.1 Comparison of StellarPGx with other callers based on *CYP2D6* diplotyping

We used published consensus *CYP2D6* diplotypes for the 109 WGS samples corresponding to the GeT-RM reference materials as the truth diplotype calls (Gaedigk, Turner, et al., 2019) except for sample NA18519. GeT-RM reported the consensus *CYP2D6* call for NA18519 as **1/*29*. However, this sample's reference diplotype was based on targeted genotype panel results only (Gaedigk, Turner, et al., 2019), thus probably missing the **106* defining variant (3878G>A), which appears in the WGS data. Therefore, we considered *CYP2D6*29/*106* as the truth call for NA18519 as did the Cyrius authors (Chen et al., 2021). The assays used to generate the GeT-RM truth calls are described in detail by Gaedigk, Turner, et al. (2019).

StellarPGx's *CYP2D6* star allele calls were highly concordant (99%) with the GeT-RM reference materials and comparable to Cyrius calls (Table 4.2). Notably, StellarPGx was able to correctly genotype sample NA19908 (*CYP2D6*1/*46* — normal metaboliser phenotype) as the pipeline queried multiple suballele combinations unlike Cyrius which ambiguously reported *CYP2D6*1/*46* or **1/*43* (indeterminate phenotype) for this sample. In comparison, Aldy and Stargazer diplotype calls had lower concordance with GeT-RM calls (Table 4.3).

Furthermore, StellarPGx's phenotype assignments were 100% concordant with those corresponding to the GeT-RM reference materials. Cyrius had comparable phenotype concordance (99%), while Aldy (92%) and Stargazer (90%) had lower phenotype assignment accuracy. Activity scores and metaboliser phenotypes were assigned according to consensus recommendations by Caudle et al. (2020). Detailed diplotype calls and corresponding phenotype assignments for each sample by each tool are provided in Table S4.1 on page 102.

Sample NA18545 (*CYP2D6*5/*36x2+*10x2*) was particularly challenging for all four tools as the **5* allele could not be detected in the presence of the *CYP2D6*36x2+*10x2* allele on the other chromosome. This sample was not genotyped by Stargazer while StellarPGx, Cyrius and Aldy genotyped it as *CYP2D6*36+*10/*36+*10*. Both diplotypes, however, translate to an

intermediate metaboliser phenotype when applying the Clinical Pharmacogenetics Implementation Consortium (CPIC) recommended translation method (Caudle et al., 2020). StellarPGx consistently called the other diplotypes comprising *CYP2D6* hybrid rearrangements (Figure 4.2).

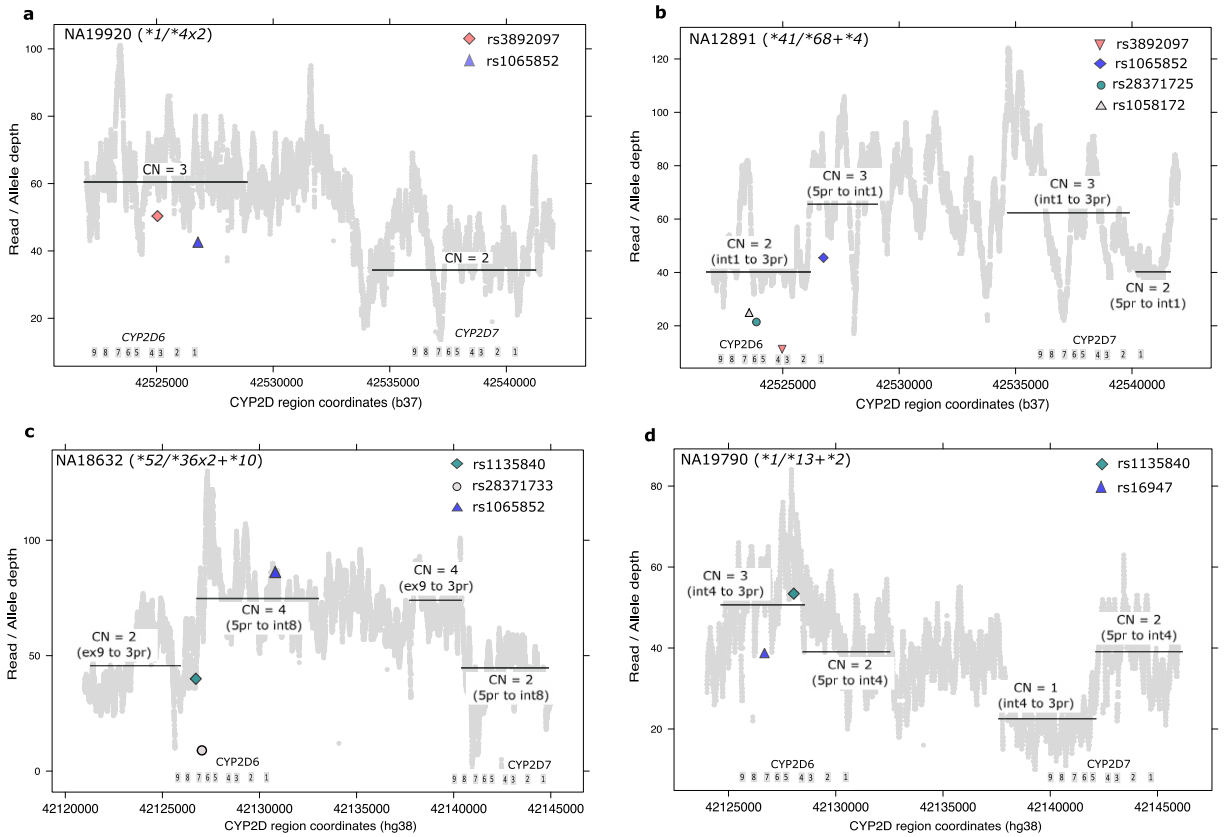


Figure 4.2: Examples of *CYP2D6* structural variants detected by StellarPGx in the WGS validation dataset corresponding to the GeT-RM reference materials. (**a**) Shows the *CYP2D* region coverage in an admixed American sample NA19920 with the *CYP2D6*1/*4×2* diplotype. Note the higher *CYP2D6* copy number (CN = 3) compared with *CYP2D7* (CN = 2) and the comparable allele depth for the core variants (rs3892097 and rs1065852). (**b**) Shows a European sample NA12891 with the *CYP2D6*41/*68+*4* diplotype. Note the hybrid read distribution due to the **68* allele and the intron 1 switch region. Also, the allele depth for rs1065852 is much higher than that for rs3892097. (**c**) Shows an East Asian sample NA18632 with the complex *CYP2D6*52/*36×2+*10* diplotype. Note the exon 9 conversion showed by diminished coverage for the *CYP2D6* exon 9 and elevated coverage for CYP2D7 exon 9. (**d**) Shows sample NA19790 with the complex *CYP2D6*1/*13+*2* diplotype. Note the intron 4 switch region for the *CYP2D6*13* hybrid allele. CYP, cytochrome P450; ex, exon; int, intron; WGS, whole genome sequence; 3pr, 3-prime region; 5pr, 5-prime region

.

Notably, some indel positions in complex gene regions had to be shifted in our allele databases to call alleles such as *CYP2D6*40* and *CYP2D6*21*. This shift in indel position is also implemented by the other allele callers.

4.4.2 StellarPGx's performance on a larger *CYP2D6* diplotype sample space

Given that the number of CYP2D6 alleles represented in the GeT-RM reference materials was rather limiting, we further tested the ability of StellarPGx's combinatorial diplotype assignment to resolve 35,778 diplotype simulations. This includes all combinations of 267 SNV-defined haplotypes covering *CYP2D6*1.001–*140.001* (see Table S4.2) from PharmVar v4.2.4 corresponding to GRCh37. StellarPGx correctly resolved 35,764 diplotypes and reported ambiguous calls for 14 cases. Examples of the ambiguous cases encountered include *CYP2D6*132.001/*9.001* (called as *CYP2D6*10/*115* or **132/*9*), *CYP2D6*1.001/*6.003* (called as *CYP2D6*1/*6* or **39/*6*) and *CYP2D6*2.004/*6.002* (called as *CYP2D6*2/*6* or **34/*6*). See Appendix D — page 226, and Table S4.2 for more details on the simulation analysis.

4.4.3 StellarPGx's performance on 11 additional CYP genes

StellarPGx's concordance with GeT-RM calls for *CYP2A6*, *CYP2B6*, *CYP2C19*, *CYP2C9*, *CYP2C8*, *CYP3A4*, *CYP3A5*, *CYP2A1*, *CYP1A2*, *CYP2E1* and *CYP4F2* diplotypes is summarised in Table 4.4 on page 98. Briefly, StellarPGx had 100% concordance with GeT-RM diplotype calls (Pratt et al., 2016) for *CYP1A2*, *CYP3A4*, *CYP3A5* and two *CYP2C* cluster genes i.e. *CYP2C9* and *CYP2C8*. We also obtained highly comparable diplotypes with GeT-RM calls for *CYP4F2* (96%), *CYP2C19* (97%) and *CYP2A6* (88%). On the other hand, genotyping *CYP1A1*, *CYP2E1*, and *CYP2B6* was more challenging as the StellarPGx concordance with GeT-RM calls with respect to these genes was 78%, 63%, and 66% respectively. For a detailed explanation of the discrepancies in the called diplotypes for *CYP1A1*, *CYP2E1*, and *CYP2B6*, see subsection 4.5. Detailed star allele calls for each gene/sample are provided in Table S4.3 and Table S4.4 on pages 106 – 109.

Of note, we identified samples with possible novel alleles for some CYP genes. For example, samples NA07056, NA07029 and NA18564 were genotyped as *CYP4F2*3/*3* by GeT-RM but StellarPGx flagged that a possible novel *CYP4F2* allele was present as rs2108622 (*CYP4F2*3*) and rs3093105 (*CYP4F2*2*) were deduced to be on the same chromosome in each of these samples.

4.4.4 Run time comparisons

We next compared the run time of StellarPGx against Aldy, Stargazer and Cyrius with regard to *CYP2D6* star allele calling while employing 20GB memory and 10 CPUs per task on up to 4 identical cluster nodes (Intel Xeon E3-12xx v2 model — 2GHz) running CentOS 7.5. 49 For a single WGS sample run, Aldy was the fastest tool with a run time of 7s while StellarPGx (58s), Stargazer (91s) and Cyrius (140s) were significantly slower. However, for a multi-sample run

Table 4.4: Summary of the diplotype concordance for StellarPGx with respect to 11 additional CYP genes in comparison to the GeT-RM reference materials. The sample number ranges from 67-75 (all aligned to build 37) as not all the 109 aforementioned WGS samples had publicly available truth calls from GeT-RM.

Gene	Number of samples	Concordant samples	Concordance (%)
CYP2C8	74	74	100
CYP2C9	74	74	100
CYP3A4	74	74	100
CYP3A5	74	74	100
CYP1A2	67	67	100
CYP2C19	74	72	97
CYP4F2	67	64	96
CYP2A6	67	59	88
CYP1A1	67	52	78
CYP2B6	74	49	66
CYP2E1	67	42	63

(75 WGS samples aligned to build 37), StellarPGx had the fastest run time (4min 7s) followed by Aldy (6min) and Stargazer (9min 51s) while Cyrius was significantly slower (57min 49s when using a single manifest file). StellarPGx was able to obtain the fastest run time for the multi-sample run owing to the efficient parallelisation facilitated by its Nextflow implementation.

4.5 Discussion

Accurately calling star alleles in CYP genes is integral to genotype-guided therapy for a wide array of drugs, and characterising pharmacogenomic variation at the population level. Reproducible bioinformatics workflows are crucial in tackling the challenging task of genotyping this class of pharmacogenes using targeted and/or WGS data from the widely used short-read NGS technologies. Existing tools notably Astrolabe, Aldy, Stargazer and Cyrius have provided short-read-based genotyping solutions for CYP and/or other pharmacogenes in recent years. However, more tools with diverse approaches are still needed especially to support high confidence ensemble CYP genotyping that we propose in our previous benchmarking study (see Chapter 3). As long read technologies such as PacBio (Buermans et al., 2017; Qiao et al., 2016) and Oxford Nanopore (Ammar et al., 2015; Liau et al., 2019) become more widely used, adapting star allele callers to the analysis of targeted and/or WGS long-read data will presumably be easier, but nonetheless important.

We have described StellarPGx, a Nextflow pipeline for determining diplotypes in CYP genes from short-read HTS data, and demonstrated its analytic accuracy using *CYP2D6* as a model CYP gene. StellarPGx uses a hybrid approach involving genome graph-based variant calling,

read coverage information from the original linear alignments and combinatorial star allele assignment. StellarPGx's *CYP2D6* allele calling performance is comparable to the recently developed Cyrius pipeline, and is superior to Aldy and Stargazer based on the analysis of WGS data corresponding to the GeT-RM reference materials.

Furthermore, StellarPGx's phenotype assignments were more concordant with the GeT-RM reference materials compared to any of the other allele callers (Table S4.1). The effect of genotyping accuracy on phenotype assignment, however, depends on the sample diplotype. For example, sample NA18565 (*CYP2D6*10/*36x2*) was correctly assigned an intermediate metaboliser phenotype by all four allele callers, even though Aldy and Stargazer incorrectly called it as *CYP2D6*10/*36+*10*. Conversely, sample NA18524 (*CYP2D6*1/*36x2+*10*, normal metaboliser) was incorrectly assigned an intermediate metaboliser phenotype by Stargazer, which genotyped it as (*CYP2D6*36+10/ *36+*10*).

Though comparable in genotyping performance, StellarPGx offers notable advantages over Cyrius. Firstly, StellarPGx is able to call diplotypes in up to 12 CYP genes while Cyrius is currently only available for *CYP2D6*. Secondly, owing to its implementation in Nextflow and containerisation with both Docker and Singularity, StellarPGx installs easily across various computational platforms and affords high reproducibility and scalability. In the same vein, StellarPGx calls star alleles in a shorter time compared to Cyrius especially for large cohorts. In addition, users can easily perform modifications to the StellarPGx pipeline by including custom Nextflow processes for performing different steps. One such modification would be to run a tool like Aldy in a parallel process so as to obtain consensus genotype calls for the same CYP gene(s) in multiple samples. Furthermore, the use of pangenome-based variant calling in the initial process of the StellarPGx pipeline enhances allele calling confidence. In addition, StellarPGx is able to detect possible occurrences of novel alleles by flagging novel combinations of existing core SNVs and also interrogating whether novel core SNVs (especially missense, frameshift, stop loss and stop gain variants) are present in a sample.

As shown in our simulations, StellarPGx is currently not able to resolve diplotypes with exactly similar variant combinations (see Appendix D), thus resulting in ambiguous diplotype assignments for such cases. This limitation is also common to other allele callers including Astrolabe (Twist et al., 2016), Aldy (Numanagić et al., 2018) and Cyrius (Chen et al., 2021). Future updates to the StellarPGx database with alleles upgraded from limited to moderate or definitive level of evidence by PharmVar will likely mitigate this challenge. Another main limitation is having to perform structural variant detection by complementing GraphTyper's structural variant genotypes with coverage information from original linear alignments. Even though genome graph-based variant detection has proved to be excellent for genotyping SNVs, full gene deletion and full gene duplications, we observed a number of false positive and false

negative genotypes for the hybrid rearrangements in the model gene (*CYP2D6*) and challenges in including them in the pangenome construction, hence resorting to the hybrid allele calling approach. Additionally, the realignment by GraphTyper to the pangenome is done for reads already aligned to a genomic region of interest in the original linear reference alignment which could introduce some reference-based alignment bias.

Further, StellarPGx's diplotype concordance was lower for *CYP1A1*, *CYP2E1*, and *CYP2B6* compared to other supported CYP genes. For *CYP2B6* in particular, StellarPGx had a high number of ambiguous diplotype calls (20 samples) with respect to the dataset used for our validation. For example, samples HG00436, NA06991, NA07029, NA07055 and NA10847 (genotyped as *CYP2B6*1/*6* by GeT-RM as **6* is considerably more frequent than *4) were genotyped as *CYP2B6*1/*6* or **4/*9* by StellarPGx since both diplotypes are equally likely given the allele defining variant combinations in the samples. For *CYP1A1* and *CYP2E1*, one possible explanation for the inconsistent diplotype calls is that the number of variants tested by some of the genotyping panels used by GeT-RM was rather limiting. For example, sample NA18980 was called as *CYP1A1*1/*2* by GeT-RM, while StellarPGx called it as *CYP1A1*13/*2B* since we were able to detect the *CYP1A1*13* defining variant rs4646422. Notably, for *CYP2E1*, only two panels were used for generating consensus reference *CYP2E1* genotypes by GeT-RM. Indeed most of the discrepant *CYP2E1* diplotypes between StellarPGx and GeT-RM arose in samples that did not have definitive consensus calls by GeT-RM. For example, samples HG00436 and NA18980 were homozygous for both *CYP2E1*5B* (rs3813867, rs2031920) and *CYP2E1*7* (rs2070673) defining variants. GeT-RM genotyped them as *CYP2E1* (**7/*7*), indicating uncertainty in the final diplotype, while StellarPGx had 'no calls' for these samples and indicated possible presence of novel alleles. Similarly, sample NA18992 was genotyped by GeT-RM as *CYP2E1*(**7*)*/*7* while StellarPGx detected the potentially function-altering rs199856651 in this sample. Furthermore, possible novel alleles were also identified for *CYP4F2* and *CYP2A6* in the validation dataset (see Table S4.3 and Table S4.4). These findings indicate that more comprehensively-characterised reference materials are needed for all core CYP genes as has been undertaken for *CYP2D6* recently (Gaedigk et al., 2019). Additionally, updates to PharmVar's star allele definitions for genes such as *CYP2E1* and *CYP2A6* will allow for allele definitions to better reflect variation in their loci.

Future developments of StellarPGx include adding support for more CYP genes as well as related pharmacogenes such as *POR*. We also intend to make StellarPGx a fully pangenome-based star allele calling pipeline as structural variant genotyping using genome graph approaches improves. Additionally, adding support for long-read sequence data and targeted NGS data from panels such as PGRNseq (Gordon et al., 2016) would enhance the utility of StellarPGx. However, analysing whole exome sequence data is still a major challenge as intronic SNVs and copy number change-points are also important for StellarPGx's star allele calling.

For short-read data, we recommend running StellarPGx using high coverage (at least 25x) WGS data especially when genotyping *CYP2D6*, *CYP2A6*, *CYP2B6* and *CYP2E1*. Further efforts aim to keep the automated phenotype prediction up to date as activity score systems and phenotype assignment recommendations are refined by clinical pharmacogenomics consortia.

We plan to continue enhancing StellarPGx's ability to detect novel alleles as well as complex population-specific alleles as community usage increases especially from studies using NGS data from under-represented populations. StellarPGx's allele databases will also be updated on a regular basis to match the updates to reference databases such as PharmVar (Gaedigk, Ingelman-Sundberg, et al., 2018), especially as comprehensive allele characterisation of genes such as *CYP2E1* and *CYP2A6* becomes available.

We recommend running StellarPGx in a multi-node cluster-environment using the pre-built Singularity image. However, for single sample cases, the pipeline can also be run on a personal computer (MacOS or Linux) preferably enhanced with external disk storage. StellarPGx is freely available from `https://github.com/SBIMB/StellarPGx` under the MIT licence.

4.6 Acknowledgements

We acknowledge the Centers for Disease Control and Prevention's (CDC) Genetic Testing Reference Material Coordination Program (GeT-RM) for providing CYP2D6 reference materials for validation of pharmacogene genotyping approaches. We also acknowledge Illumina for generating 75 of the WGS samples (including the Polaris pharmacogenomics cohort) used in the validation of StellarPGx. Samples NA19790, NA10860 and HG00423 were generated from Coriell cell lines at the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository. Thirty one of the high coverage 1000 Genomes project datasets were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

4.7 Author contributions

D.T., B.I.D., G.E.B.W., A.S., J.dR., Z.L., and S.H. wrote the manuscript; D.T., B.I.D., G.E.B.W., J.dR., Z.L., and S.H. Designed the research; D.T. and S.H. performed the research; D.T. analysed the data.

Table S4.1: *CYP2D6* diplotype calls and phenotype assignments by StellarPGx in comparison to the existing allele callers

#	Sample ID	GeT-RM			StellarPGx			Cyrius			Aldy			Stargazer		
		Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype
1	HG00111	*3/*3	0	PM	*3/*3	0	PM	*3/*3	0	PM	*3/*3	0	PM	*3/*3	0	PM
2	HG00337	*22/*2x2	IND	IND	*22/*2x2	IND	IND	*22/*2x2	IND	IND	*22/*2x2	IND	IND	*22/*83+*2	IND	IND
3	HG00373	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM
4	HG00421	*2/*10x2	1.5	NM	*2/*10x2	1.5	NM	*2/*10x2	1.5	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM
5	HG00423	*10/*10x2	0.75	IM	*10/*10x2	0.75	IM	*10/*10x2	0.75	IM	*10/*10x2	0.75	IM	*10/*36+*10	0.5	IM
6	HG00463	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM
7	HG01086	*1/*31	1	IM	*1/*31	1	IM	*1/*31	1	IM	*1/*31	1	IM	*31/*122	IND	IND
8	HG01094	*1/*31	1	IM	*1/*31	1	IM	*1/*31	1	IM	*1/*31	1	IM	*1/*31	1	IM
9	HG01108	*106/*2	IND	IND	*106/*2	IND	IND	*106/*2	IND	IND	*106/*2	IND	IND	*2/*106	IND	IND
10	HG01680	*28/*59	IND	IND	*28/*59	IND	IND	*28/*59	IND	IND	*28/*59	IND	IND	*28/*59	IND	IND
11	HG02373	*14/*36+*10	0.75	IM	*14/*36+*10	0.75	IM	*14/*36+*10	0.75	IM	*14/*36+*10	0.75	IM	*14/*36+*10	0.75	IM
12	HG03225	*5/*56	0	PM	*5/*56	0	PM	*5/*56	0	PM	*5/*56	0	PM	*5/*56	0	PM
13	HG03246	*5/*43	IND	IND	*5/*43	IND	IND	*5/*43	IND	IND	NG	NG	NG	*5/*43	IND	IND
14	HG03259	*5/*106	IND	IND	*5/*106	IND	IND	*5/*106	IND	IND	*5/*106	IND	IND	*5/*106	IND	IND
15	HG03619	*113/*2	IND	IND	*113/*2	IND	IND	*113/*2	IND	IND	*113/*2	IND	IND	*1/*2	IND	IND
16	HG03643	*2/*7	1	IM	*2/*7	1	IM	*2/*7	1	IM	*2/*7	1	IM	*2/*7	1	IM
17	HG03703	*1/*99	1	IM	*1/*99	1	IM	*1/*99	1	IM	*1/*10	1.25	NM	*1/*99	1	IM
18	HG03780	*1/*112	IND	IND	*1/*112	IND	IND	*1/*112	IND	IND	*1/*1-like	IND	IND	*1/*112	IND	IND
19	HG03781	*2/*99	1	IM	*2/*99	1	IM	*2/*99	1	IM	*2/*10	1.25	IM	*2/*99	1	IM
20	HG03882	*1/*112	IND	IND	*1/*112	IND	IND	*1/*112	IND	IND	*1/*1-like + *61	IND	IND	*1/*112	IND	IND
21	HG04206	*113/*2	IND	IND	*113/*2	IND	IND	*113/*2	IND	IND	*1/*2	2	NM	*2/*113	IND	IND
22	NA06989	*9/*9	1	IM	*9/*9	1	IM	*9/*9-like	1	IM	*9/*9	1	IM	*9/*9	1	IM
23	NA10860	*1/*4x2	1	IM	*1/*4x2	1	IM	*1/*4N+*4	1	IM	*1/*4x2	1	IM	*1/*4N+*4	1	IM
24	NA12154	*33/*68+*4	1	IM	*33/*68+*4	1	IM	*33/*68+*4	1	IM	*33/*68+*4	1	IM	*33x2/*68+*4	2	NM
25	NA18545	*5/*36x2+*10x2	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	NG	NG	NG
26	NA18632	*52/*36x2+*10	IND	IND	*52/*36x2+*10	IND	IND	*52/*36x2+*10	IND	IND	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM
27	NA18642	*1+*90/*36+*10	IND	IND	*1+*90/*36+*10	IND	IND	*1+*90/*36+*10	IND	IND	*1+*90/*36+*10	IND	IND	*1x2/*36+*10	2.25	NM
28	NA19317	*5/*5	0	PM	*5/*5	0	PM	*5/*5	0	PM	NG	NG	NG	*5/*5	0	PM
29	NA19777	*1/*82	IND	IND	*1/*82	IND	IND	*1/*82	IND	IND	*1/*74-like or 1-like/*74	IND	IND	*1/*1	2	NM
30	NA19790	*1/*13+*2	2	NM	*1/*13+*2	2	NM	*1/*13+*2	2	NM	*1/*78+*2	2	NM	*2x2/*13B	2	NM
31	NA20289	*11/*6	0	PM	*11/*6	0	PM	*11/*6	0	PM	*11/*6	0	PM	*1/*6	1	IM
32	NA20803	*2/*22	IND	IND	*2/*22	IND	IND	*2/*22	IND	IND	*2/*22	IND	IND	*2/*22	IND	IND
33	NA20875	*1/*111	IND	IND	*1/*111	IND	IND	*1/*111	IND	IND	*1/*111	IND	IND	*1/*111	IND	IND
34	NA21105	*111/*3	IND	IND	*111/*3	IND	IND	*111/*3	IND	IND	*111/*2-like or *2-like/*3	IND	IND	*3/*111	IND	IND
35	NA18484	*1/*17	1.5	NM	*1/*17	1.5	NM	*1/*17	1.5	NM	*61-like/*67 or *34/*78	IND	IND	*1/*17	1.5	NM

Table S4.1 – continued

#	Sample ID	GeT-RM			StellarPGx			Cyrius			Aldy			Stargazer		
		Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype
36	NA18509	*2/*17	1.5	NM	*17/*2	1.5	NM	*2/*17	1.5	NM	*17/*2	1.5	NM	*2/*17	1.5	NM
37	NA18518	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM
38	NA18519	*1/*29	1.5	NM	*106/*29	1.5	NM	*106/*29	1.5	NM	*106/*29	1.5	NM	*29/*106	1.5	NM
39	NA18855	*1/*5	1	IM	*5/*1	1	IM	*1/*5	1	IM	*1/*5	1	IM	*1/*5	1	IM
40	NA18861	*5/*29	0.5	IM	*29/*5	0.5	IM	*5/*29	0.5	IM	*29/*5	0.5	IM	*5/*29	0.5	IM
41	NA18868	*2/*5	1	IM	*2/*5	1	IM	*2/*5	1	IM	*2/*5	1	IM	*2/*5	1	IM
42	NA19095	*1/*29	1.5	NM	*1/*29	1.5	NM	*1/*29	1.5	NM	*1/*29	1.5	NM	*1/*29	1.5	NM
43	NA19109	*2x2/*29	2.5	UM	*29/*2x2	2.5	UM	*2x2/*29	2.5	UM	*2x2/*29	2.5	UM	*2x2/*29	2.5	UM
44	NA19122	*2/*17	1.5	NM	*17/*2	1.5	NM	*2/*17	1.5	NM	*17/*2	1.5	NM	*2/*17	1.5	NM
45	NA19143	*2(*45)/*10	1.25	NM	*10/*45	1.25	NM	*10/*45	1.25	NM	*10/*45	1.25	NM	*2(*45)/*10	1.25	NM
46	NA19147	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM	*17/*29	1	IM	*17/*39	1.5	NM
47	NA19174	*4/*40	0	PM	*4/*40	0	PM	*4/*40	0	PM	*4/*40	0	PM	*4/*40	0	PM
48	NA19176	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM
49	NA19178	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM
50	NA19207	*2x2/*10	2.25	NM	*10/*2x2	2.25	NM	*2x2/*10	2.25	NM	*2x2/*10	2.25	NM	*10/*2x2	2.25	NM
51	NA19213	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM
52	NA19226	*2/*2x2	3	UM	*2/*2x2	3	UM	*2/*2x2	3	UM	*2/*2x2	3	UM	*2/*2x2	3	UM
53	NA19238	*1/*17	1.5	NM	*1/*17	1.5	NM	*1/*17	1.5	NM	*1/*17	1.5	NM	*1/*17	1.5	NM
54	NA19239	*15/*17	0.5	IM	*15/*17	0.5	IM	*15/*17	0.5	IM	*15/*17	0.5	IM	*1/*15	1	IM
55	NA19240	*15/*17	0.5	IM	*15/*17	0.5	IM	*15/*17	0.5	IM	*15/*17	0.5	IM	*1/*15	1	IM
56	HG01190	*5/*68+*4	0	UM	*5/*68+*4	0	UM	*5/*68+*4	0	UM	*4/*68	0	UM	*5/*68+*4	0	UM
57	NA19789	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*61	IND	NM	*1/*127	IND	IND
58	NA19819	*2/*4x2	1	IM	*2/*4x2	1	IM	*2/*4x2	1	IM	*2/*4x2	1	IM	*2/*4x2	1	IM
59	NA19908	*1/*46	2	NM	*1/*46	2	NM	*1/*46 or *43/*45	IND	IND	*1/*46 or *43/*45	IND	IND	*1/*46	2	NM
60	NA19917	*1/*40	1	IM	*1/*40	1	IM	*1/*40	1	IM	*1/*40	1	IM	*1/*40	1	IM
61	NA19920	*1/*4x2	1	IM	*1/*4x2	1	IM	*1/*4x2	1	IM	*1/*4x2	1	IM	*1/*4x2	1	IM
62	NA20296	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*63/*67 or *63/*78	IND	NM	*1/*2	2	NM
63	HG00436	*2x2/*71	IND	IND	*2x2/*71	IND	IND	*2x2/*71	IND	IND	*2x2/*71	IND	IND	*71/*2x2	IND	IND
64	HG00589	*1/*21	1	IM	*1/*21	1	IM	*1/*21	1	IM	*1/*21	1	IM	*1/*21	1	IM
65	NA18524	*1/*36x2+*10	1.25	NM	*1/*36x2+*10	1.25	NM	*1/*36x2+*10	1.25	NM	*1+*36/*36+*10	1.25	NM	*36+*10/*36+*10	0.5	IM
66	NA18526	*1/*36x2+*10	1.25	NM	*1/*36x2+*10	1.25	NM	*1/*36x2+*10	1.25	NM	*1+*36/*36+*10	1.25	NM	*36+*10/*36+*10	0.5	IM
67	NA18540	(*36+)*10/*41	0.75	IM	*41/*36+*10	0.75	IM	*41/*36x2+*10	0.75	IM	*36+*10/*61+*69	IND	IM	*36+*10/*36+*10	0.5	IM
68	NA18544	*10/*41	0.75	IM	*10/*41	0.75	IM	*10/*41	0.75	IM	*10/*41	0.75	IM	*10/*41	0.75	IM
69	NA18552	*1/*14	1.5	NM	*1/*14	1.5	NM	*1/*14	1.5	NM	*1/*14	1.5	NM	*1/*14	1.5	NM
70	NA18564	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM
71	NA18565	*10/*36x2	0.25	IM	*10/*36x2	0.25	IM	*10/*36x2	0.25	IM	*10/*36+*10	0.5	IM	*10/*36+*10	0.5	IM
72	NA18617	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM	*36+*10/*36+*10	0.5	IM
73	NA18942	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM
74	NA18952	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM	*2/*2	2	NM

Table S4.1 – continued

#	Sample ID	GeT-RM			StellarPGx			Cyrius			Aldy			Stargazer		
		Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype	Diplotype calls	AS	Phenotype
75	NA18959	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM
76	NA18966	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM
77	NA18973	*1/*21	1	IM	*1/*21	1	IM	*1/*21	1	IM	*1/*21	1	IM	*1/*21	1	IM
78	NA18980	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM	*2/*36+*10	1.25	NM
79	NA18992	*1/*5	1	IM	*5/*1	1	IM	*1/*5	1	IM	*1/*5	1	IM	*1/*5	1	IM
80	NA19003	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM
81	NA19007	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM
82	HG00276	*4/*5	0	PM	*4/*5	0	PM	*4/*5	0	PM	*4/*5	0	PM	*4/*5	0	PM
83	NA06991	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM
84	NA07000	*9/*2(*35)	1.5	NM	*35/*9	1.5	NM	*9/*35	1.5	NM	*35/*9	1.5	NM	*9/*35	1.5	NM
85	NA07019	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM
86	NA07029	*1/*35	2	NM	*1/*35	2	NM	*1/*35	2	NM	*1/*35	2	NM	*1/*35	2	NM
87	NA07055	*4/*4	0	PM	*4/*4	0	PM	*4/*4	0	PM	*4/*4	0	PM	*4/*4	0	PM
88	NA07056	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM
89	NA07348	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM
90	NA07357	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM	*1/*6	1	IM
91	NA10831	*5/*4	0	PM	*4/*5	0	PM	*5/*4	0	PM	*4/*5	0	PM	*5/*4	0	PM
92	NA10847	*1/*41	1.5	NM	*1/*41	1.5	NM	*1/*41	1.5	NM	*1/*41	1.5	NM	*1/*41	1.5	NM
93	NA10851	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM
94	NA10854	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM
95	NA11832	*1/(*68)+*4	1	IM	*1/*68+*4	1	IM	*1/*68+*4	1	IM	*1/*68+*4	1	IM	*1/*68+*4	1	IM
96	NA11839	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM	*1/*2	2	NM
97	NA11993	*1/*9	1.5	NM	*1/*9	1.5	NM	*1/*9	1.5	NM	*1/*9	1.5	NM	*1/*9	1.5	NM
98	NA12003	*4/*35	1	IM	*35/*4	1	IM	*4/*35	1	IM	*35/*4	1	IM	*4/*35	1	IM
99	NA12006	*4/*41	0.5	IM	*4/*41	0.5	IM	*4/*41	0.5	IM	*4/*41	0.5	IM	*4/*41	0.5	IM
100	NA12145	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM
101	NA12156	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM	*1/*4	1	IM
102	NA12717	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM	*1/*1	2	NM
103	NA12813	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM	*2/*4	1	IM
104	NA12873	*1/*5	1	IM	*1/*5	1	IM	*1/*5	1	IM	*NG*	NG	NG	*1/*5	1	IM
105	NA20509	*4/*35	1	IM	*35/*4	1	IM	*4/*35	1	IM	*35/*4	1	IM	*4/*35	1	IM
106	NA21781	*2x2/*68+*4	2	NM	*2x2/*68+*4	2	NM	*2x2/*68+*4	2	NM	*2x2/*68+*4	2	NM	*2x2/*68+*4	2	NM
107	NA12892	*2/*3	1	IM	*2/*3	1	IM	*2/*3	1	IM	*2/*3	1	IM	*2/*3	1	IM
108	NA12891	*41/*68+*4	0.5	IM	*41/*68+*4	0.5	IM	*41/*68+*4	0.5	IM	*41/*68+*4	0.5	IM	*41/*68+*4	0.5	IM
109	NA12878	*3/*68+*4	0	PM	*3/*68+*4	0	PM	*3/*68+*4	0	PM	*3/*68+*4	0	PM	*3/*68+*4	0	PM

NG, Not genotyped; AS, Activity score; PM, Poor Metaboliser (AS = 0); IM, Intermediate metaboliser (0 < AS < 1.25); NM, Normal Metaboliser (1.25 $\leq$ AS $\leq$ 2.25); UM, Ultrarapid metaboliser (AS > 2.25)

Table S4.2: *CYP2D6* diplotypes called by StellarPGx from 35,778 simulated samples. This table is too large to be reprinted in this Thesis. The full table can be found at `https://ascpt.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fcpt.2173&file=cpt2173-sup-0005-TableS4.xlsx`

Table S4.3: StellarPGx's star allele calls for *CYP1A1*, *CYP1A2*, *CYP2A6*, *CYP2B6*, *CYP2C8* and *CYP2C9* compared to the GeT-RM reference materials

Coriell #	*CYP1A1*		*CYP1A2*		*CYP2A6*		*CYP2B6*		*CYP2C8*		*CYP2C9*	
	GeT-RM	StellarPGx	GeT-RM	StellarPGx	GeT-RM	StellarPGx	GeT-RM	StellarPGx	GeT-RM	StellarPGx	GeT-RM	StellarPGx
HG00276	**1 / *1*	**1/*1*	**1A/*1F*	**1B/*1M*	**1/*1*	**1/*1*	**2/(*4)*	**2/*4*	**1/*3*	**1/*3*	**1/*2*	**1/*2*
HG00436	**1 / *1*	*[*1/*13]*	**1A/*1F*	**1B/*1M*	**9/*9*	**4/*9*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
HG00589	**2 / *2*	**2B/*2B*	**1A/*1L,*1C/*1F*	**1B/*1L*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
HG01190	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*1*	**1/*1*	**1(*5)/*1(*27)*	**1/*5*	**1/*3*	**1/*3*	**1/*2*	**1/*61*
NA06991	**1 / *1*	**1/*1*	**1F/*1F*	**1M/*1M*	**1/*1*	**1/*1*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA07000	*NA*	**1/*1*	*NA*	**1M/*1M*	*NA*	**1/*1*	**1/*1*	**1/*1*	**1 / *1*	**1/*1*	**1/*1*	**1/*1*
NA07019	**1 / *4*	**1/*4*	**1A/*1F*	**1B/*1M*	**1/*1*	**1/*1*	**1(*5)/*1(*22)*	**22/*5*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA07029	**1 / *1*	**1/*1*	**1A/*1F*	**1B/*1M*	**1/*1*	**1/*1*	**6/(*27)*	**1/*6 or *4/*9*	**1/*3*	**1/*3*	**1/*2*	**1/*2*
NA07055	**1 / *5*	**1/*5*	**1F/*1F*	**1M/*1M*	**1/*1*	**1/*1*	**6/(*27)*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA07056	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*1*	**1/*1*	**6/(*22)*	**22/*6*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA07348	**1 / *1*	**1/*1*	**1F/*1F*	**1M/*1M*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*4*	**1/*4*	**1/*1*	**1/*1*
NA07357	**1 / *1*	**1/*1*	**1F/*1F*	**1M/*1M*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*4*	**1/*4*	**1/*1*	**1/*1*
NA10831	**1 / *1*	**1/*1*	**1A/*1F*	**1B/*1M*	**1/*2*	**1/*2*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*2*	**1/*2*
NA10847	**1 / *1*	**1/*1*	**1F/*1F*	**1M/*1M*	**1/*1*	**1/*1*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA10851	**1 / *1*	**1/*2A*	**1F/*1F*	**1M/*1N*	**1/*2*	**1/*2*	**1/*1 (*27)*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA10854	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**3/*3*	**3/*3*	**2/*2*	**2/*2*
NA11832	**1 / *1*	**1/*1*	**1F/*1F*	**1M/*1M*	**1/*2*	**1/*2*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*3 (*18)*	**1/*3*
NA11839	**1 / *1*	**1/*1*	**1A/*1F*	**1B/*1M*	**1/*9*	**1/*9*	**1/*1 (*15)*	**1/*15*	**1/*3*	**1/*3*	**2/*3 (*18)*	**2/*3*
NA11993	**1 / *1*	**1/*1*	**1F/*1F*	**1M/*1M*	**9/*17*	*PNA*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA12003	**1 / *1*	**1/*1*	**1A/*1F*	**1M/*1M*	**1/(*8)*	**1/*1*	**6/*6*	**6/*6*	**1/*3*	**1/*3*	**1/*2*	**1/*2*
NA12006	**1 / *4*	**1/*4*	**1F/*1F*	**1M/*1M*	**1/*1*	**1/*1*	**1/*1 (*5)*	**1/*5*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA12145	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*4*	**1/*4*	**1/*1*	**1/*1*
NA12156	**1 / *2*	**1/*2B or *2A/*2C*	**1A/*1F*	**1B/*1M*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*2*	**1/*2*
NA12717	**1 / *1*	**1/*1*	**1F/*1F*	**1M/*1M*	**1/*1*	**1/*1*	**1/*7*	**1/*7*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA12813	**1 / *2*	**2A/*2B*	**1F/*1L*	**1L/*1N*	**1/*1*	**1/*1*	**1/*2*	**1/*2*	**1/*1*	**1/*1*	**1/*3 (*18)*	**1/*3*
NA12873	**1 / *1*	**1/*1*	**1A/*1F*	**1B/*1M*	**1/*9*	**1/*9*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA12878	**1 / *1*	**1/*1*	**1F/*1F*	**1M/*1M*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*3*	**1/*3*	**1/*2*	**1/*2*
NA12892	*NA*	**1/*1*	*NA*	**1M/*1M*	*NA*	**1/*1*	**1/*1*	**1/*5*	**1 / *3*	**1/*3*	**1/*2*	**1/*2*
NA18484	**1 / *1*	**1/*1*	**1/*1*	**1/*1*	**1/*9*	**1/*9*	**1/*18*	**1/*18*	**1/*4*	**1/*4*	**1/*9*	**1/*9*
NA18509	**1 / *1*	**1/*1*	**1A/*1F*	**1/*1*	**1/*17*	**1/*17*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18518	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*17*	**1/*17*	**1/*6*	**1/*6 or *4/*9*	**1/*2*	**1/*2*	**1/*1*	**1/*1*

Table S4.3 – continued

Coriell #	*CYP1A1*		*CYP1A2*		*CYP2A6*		*CYP2B6*		*CYP2C8*		*CYP2C9*	
	GeT-RM	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**
NA18519	**1 / *1*	**1/*1*	**1/*1*	**1B/*1L*	**1/*1*	**1/*1*	**1/*6*	*[*17/*6]*	**1/*2*	**1/*2*	**1/*5*	**1/*5*
NA18524	**1 / *1*	**1/*1*	**1A/*1F*	**1B/*1N*	**1/*1*	**1/*1*	**1/*1 (*27)*	**1/*1*	**1/*1*	**1/*1*	**1/*3*	**1/*3*
NA18526	**1 / *1*	*[*11/*2A]*	**1A/*1L,*1C/*1F*	**1B/*1L*	**1/*9*	**7/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18540	**1 / *2*	*[*13/*2B]*	**1L/*1L*	**1L/*1L*	**1/*1*	**1/*1*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18544	**1 / *2*	**2A/*2B*	**1F/*1L*	**1L/*1N*	**1/*1*	**19/*1B10 or *19/*1B8*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18552	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**9/*9*	**9/*9*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18564	**1 / *1*	*[*1/*13]*	**1A/*1A*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18565	**2 / *2*	**2B/*2B*	**1F/*1F*	**1M/*1P*	**9/*9*	**4/*15*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18617	*NA*	**1/*1*	*NA*	**1/*1*	*NA*	**4/*1*	**1/*1*	**1/*4*	**1 / *1*	**1/*1*	**1/*1*	**1/*1*
NA18855	**1 / *1*	**1/*1*	**1/*1*	**1B/*1L*	**1/(*8)*	**1/*1*	**6/*6*	**6/*6*	**1/*1*	**1/*1*	**1/*9*	**1/*9*
NA18861	**1 / *1*	**1/*1*	**1A/*1L,*1C/*1F*		**1/*1*	**25/*1x2*	**1/*18*	**1/*18*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18868	**1 / *1*	**1/*2A*	**1A/*1L,*1C/*1F*	**1B/*1L*	**1/*17*	**1/*17*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18942	**1 / *1*	*[*13/*13]*	**1A/*1F*	**1/*1*	**1/*1*	**4/*7*	**1/*2*	**1/*2*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18952	**1 / *2*	*[*13/*2B]*	**1A/*1L,*1C/*1F*	**1B/*1L*	**4/*4*	**4/*4*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18959	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/(*4)*	**4/*1*	**1/*1(*27)*	**1/*1*	**1/*1*	**1/*1*	**1/*3 (*18)*	**1/*3*
NA18966	**1 / *1*	**1/*1*	**1A/*1F*	**1/*1*	**1/(*4)*	**4/*1*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18973	**1 / *1*	**1/*1*	**1F/*1L*	**1L/*1N*	**4/*4*	**4/*4*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18980	**1 / *2*	*[*13/*2B]*	**1A/*1L,*1C/*1F*	**1B/*1L*	**9/*9*	**9/*9*	**6/*6*	**6/*6*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18992	**1 / *1*	*[*1/*13]*	**1A/*1A*	**1/*1*	**1/*9*	**19/*9*	**1/*6*	**1/*6*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA19003	**1 / *1*	*[*1/*13]*	**1A/*1F*	**1B/*1N*	**1/*1*	**1/*1*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA19007	**1 / *2*	*[*13/*2B]*	**1A/*1L,*1C/*1F*	**1B/*1L*	**1/(*4)*	**4/*1*	**1/*1*	*[*1/*23]*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA19095	**1 / *1*	**1/*1*	**1/*1*	**1/*1*	**9/*9*	**9/*9*	**18/*18*	**18/*18*	**1/*2*	**1/*2*	**1/*1*	**1/*1*
NA19109	**1 / *1*	**1/*1*	**1A/*1F*	**1B/*1N*	**17/*20*	*PNA*	**1/*6*	**1/*6 or *4/*9*	**2/*2*	**2/*2*	**1/*1*	**1/*1*
NA19122	**1 / *1*	**1/*1*	**1A/*1L,*1C/*1F*	**1B/*1L*	**1/*1*	**1/*35*	**6/*6*	**6/*6*	**1/*1*	**1/*1*	**1/*11*	**1/*11*
NA19143	**1 / *1*	*[*1/*2A]*	**1A/*1F*	**1/*1*	**1/*1*	**1/*35*	**6/*6*	**6/*6*	**1/*1*	**1/*1*	**1/*6*	**1/*6*
NA19147	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*9*	**23/*9*	**1/*18*	**1/*18*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA19174	**1 / *1*	*[*1/*2A]*	**1A/*1L,*1C/*1F*	**1B/*1L*	**1/*9*	**24/*9*	**6/*18*	**6/*18 or *16/*9*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA19176	**1 / *1*	**1/*1*	**1/*1*	**1/*1*	**1/(*8)*	**1/*1*	**6/*6*	**6/*6*	**2/*2*	**2/*2*	**1/*1*	**1/*1*
NA19178	**1 / *1*	*[*1/*2A]*	**1L/*1L*	**1L/*1L*	**1/*20*	*PNA*	**6/*6*	*[*6/*(full_gene_del)]*	**1/*1*	**1/*1*	**5/*9*	**5/*9*
NA19207	**1 / *1*	*[*1/*2A]*	**1A/*1F*	**1/*1*	**1/*17*	**1/*17*	**1/*6*	**1/*6 or *4/*9*	**1/*2*	**1/*2*	**1/*1*	**1/*1*
NA19213	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*17*	**17/*24*	**6/*6*	**6/*6*	**1/*1*	**1/*1*	**1/*6*	**1/*6*
NA19226	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*17*	**1/*17*	**18/(*20)*	**18/*20*	**1/*1*	**1/*1*	**1/*8*	**1/*8*
NA19238	*NA*	**1/*1*	*NA*	*PNA*	*NA*	**1/*9*	**1/*6*	**1/*6 or *4/*9*	**1 / *1*	**1/*1*	**1/*1*	**1/*1*
NA19239	**1 / *1*	*[*1/*2A]*	**1A/*1L,*1C/*1F*	**1B/*1L*	**1/*17*	**1/*17*	**1/*6*	**1/*6 or *4/*9*	**1/*2*	**1/*2*	**1/*1*	**1/*1*

Table S4.3 – continued

Coriell #	*CYP1A1*		*CYP1A2*		*CYP2A6*		*CYP2B6*		*CYP2C8*		*CYP2C9*	
	GeT-RM	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**
NA19789	**1 / *2*	**1/*2B or *2A/*2C*	**1A/*1L,*1C/*1F*	**1B/*1L*	**1/*1*	**1/*1*	**1/*1 (*27)*	**1/*1*	**1/*3*	**1/*3*	**1/*2*	**1/*2*
NA19819	**1 / *1*	*[*1/*2A]*	**1F/*1L*	**1L/*1M*	**1/*1*	**1/*1*	**1/*1*	**1/*1*	**1/*2*	**1/*2*	**1/*1*	**1/*1*
NA19908	**1 / *2*	*[*2A/*2B]*	**1L/*1L*	**1L/*1L*	**1/*17*	**1/*17*	**1/*6*	**1/*6 or *4/*9*	**1/*1*	**1/*1*	**1/*5*	**1/*5*
NA19917	*NA*	**1/*1*	*NA*	**1B/*1N*	*NA*	**1/*1*	**6/*6*	**6/*6*	**1 / *1*	**1/*1*	**1/*1 (*18)*	**1/*1*
NA19920	**1 / *1*	**1/*1*	**1A/*1A*	**1/*1*	**1/*1*	**1/*35*	**6/*6*	**6/*6*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA20296	**1 / *1*	*[*1/*2A]*	**1L/*1L*	**1L/*1L*	**1/*1*	**1/*1*	**1/*2*	**1/*2*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA20509	**1 / *1*	*[*1/*2A]*	**1A/*1F*	**1B/*1M*	**1/*1*	**1/*1*	**1/*7*	**5/*9*	**1/*4*	**1/*4*	**1/*1*	**1/*1*
NA21781	*NA*	**1/*1*	*NA*	**1M/*1M*	*NA*	**1/*21*	**1/*1*	**1/*4*	**1 / *1*	**1/*1*	**1/*1*	**1/*1*
NA12891	*NA*	**1/*1*	*NA*	**1M/*1M*	*NA*	**1/*14*	*NA*	**1/*5*	*NA*	**1/*4*	*NA*	**1/*1*
NA19240	*NA*	**1/*1*	*NA*	**1/*1*	*NA*	**4/*9*	**1/*1*	**1/*1*	**1/*2*	**1/*2*	**1/*1*	**1/*1*

PNA, Possible novel alleles detected; NA, Truth call not publicly available; (*n/*n or *m/*m) — Ambiguous call; *n/*n — Discordant call; [*m/*m] — Discordant cases where StellarPGx identified allele defining variants that were either not identified or not tested by the genotyping panels used by GeT-RM

Notes:
- The GeT-RM calls included here correspond to the reference materials characterised by Pratt et al. (2016) Notably, some alleles have now been redefined by PharmVar as outlined below.
- *CYP2C19*38* was formerly *CYP2C19*1A* (see `https://www.pharmvar.org/gene/CYP2C19`; last accessed 14-Dec-2020);
so we do not consider cases where StellarPGx called CYP2C19*38 and GeT-RM called *CYP2C19*1* as discrepancies.
- *CYP2C19*27* has been redefined as *CYP2C19*1.006* by PharmVar (see `https://www.pharmvar.org/gene/CYP2C19`; accessed 14-Dec-2020).

Table S4.4: StellarPGx's star allele calls for *CYP2C19*, *CYP2E1*, *CYP3A4*, *CYP3A5* and *CYP4F2* compared to the GeT-RM reference materials

Coriell #	*CYP2C19*		*CYP2E1*		*CYP3A4*		*CYP3A5*		*CYP4F2*	
	GeT-RM	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**
HG00276	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*2*	**1A/*2*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
HG00436	**1/*1*	**1/*1*	*(*7 / *7)*	*PNA*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	*(*2)/*3*	**2/*3*
HG00589	**1/*1*	**1/*1*	*(*5)/ *7*	**5B/*7A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
HG01190	**1/*2*	**1/*2*	**1 / *7*	**1B/*7B*	**1/*1B*	**1B/*1G*	**1/*1*	**1/*1*	**1/*3*	**1/*3*
NA06991	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA07000	**1/*17*	**1/*17*	*NA*	**1A/*1A*	**1/*1*	**1F/*1G*	**1/*3*	**1/*3*	*NA*	*PNA*
NA07019	**1/*17*	**1/*17*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*3*	**1/*3*
NA07029	**8/*17*	*PNA*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1G*	**1/*3*	**1/*3*	**3/*3*	*PNA*
NA07055	**1/*17*	**1/*17*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1F*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA07056	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*22*	**1A/*22*	**3/*3*	**3/*3*	**3/*3*	*PNA*
NA07348	**2/*17*	**17/*2*	**1 / *7*	**1B/*7C*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA07357	**2/*17*	**17/*2*	*(*7) / *7*	*PNA*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA10831	**1/*17*	**1/*17*	**1 / *7*	**1B/*7B*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA10847	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA10851	**1/*17*	**1/*17*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA10854	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1B*	**1B/*1*	**1/*3*	**1/*3*	**1/*3*	**1/*3*
NA11832	**1/*2*	**2/*38*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA11839	**1/*1*	**1/*38*	*(*5)/ *7*	**5B/*7A*	**1/*1B*	**1B/*1*	**1/*3*	**1/*3*	**1/*3*	**1/*3*
NA11993	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1F*	**3/*3*	**3/*3*	**1/*3*	**1/*3*
NA12003	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1B*	**1B/*1*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA12006	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*3*	**1A/*3*	**3/*3*	**3/*3*	*(*2)/*3*	**1/*3*
NA12145	**2/*17*	**17/*2*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1F*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA12156	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	*(*2)/*3*	**2/*3*
NA12717	**2/*2*	**2/*2*	**1 / *1*	**1A/*1A*	**1B/*22*	**1B/*22*	**1/*3*	**1/*3*	**1/*3*	**1/*3*
NA12813	**1/*17*	**17/*38*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1G*	**3/*3*	**3/*3*	**1/*3*	**1/*3*
NA12873	**1/*17*	**1/*17*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA12878	**1/*2*	**1/*2*	*(*5)/ *7*	**5B/*7A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA12892	**1/*1*	**1/*1*	*NA*	*PNA*	**1/*22*	**1A/*22*	**3/*3*	**3/*3*	*NA*	**1/*1*
NA18484	**1 (*27)/*2*	**1/*2*	**1 / *7*	**1B/*7C*	**1B/*1B*	**1B/*1B*	**1/*7*	**1/*7*	**1/*2*	**1/*2*
NA18509	**2/*2*	**2/*2*	*(*7 / *7)*	*PNA*	**1/*1B*	**1B/*1G*	**1/*7*	**1/*7*	*(*2)/*3*	**2/*3*
NA18518	**2/*17*	**17/*2*	*(*4) / *7*	*PNA*	**1B/*1B*	**1B/*1B*	**1/*6*	**1/*6*	**1/*2*	**1/*2*
NA18519	**1/*17*	**1/*17*	**1 / *1*	**1A/*1B*	**1B/*1B*	**1B/*1B*	**1/*6*	**1/*6*	*(*2)/*3*	**2/*3*
NA18524	**1/*2*	**2/*38*	*(*7) / *7*	*PNA*	**1/*1*	**1A/*1G*	**1/*3*	**1/*3*	**1/*3*	**1/*3*
NA18526	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1G*	**1/*1*	**1/*1*	**1/*3*	**1/*3*
NA18540	**1/*2*	**1/*2*	*(*5)/ *7*	**5B/*7A*	**1/*1*	**1A/*1G*	**1/*3*	**1/*3*	**1/*3*	**1/*3*
NA18544	**1/*2*	**1/*2*	**7 / *7*	*PNA*	**1/*1*	**1A/*1A*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA18552	**1/*4*	**1/*4*	*(*5)/ *7*	*PNA*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*3*	**1/*3*
NA18564	**2/*3*	**2/*3*	*(*5)/ *7*	**5B/*7A*	**1/*1*	**1G/*1H*	**1/*1*	**1/*1*	**3/*3*	*PNA*
NA18565	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1G*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA18617	**1/*2*	**1/*2*	*NA*	*PNA*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	*NA*	**1/*1*
NA18855	**1 (*27)/*2*	**1/*2*	*(*7) / *7*	*PNA*	**1/*1B*	**1S/*1B*	**3/*6*	**3/*6*	**1/*1*	**1/*1*
NA18861	**1/*1*	**1/*1*	*(*7) / *7*	*PNA*	**1B/*1B*	**1B/*1B*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA18868	**1/*2*	**1/*2*	*(*7) / *7*	*PNA*	**1B/*1B*	**1B/*1B*	**1/*3*	**1/*3*	**1/(*2)*	**1/*1*
NA18942	**1/*1*	**1/*1*	*(*5)/ *7*	**5B/*7A*	**1/*1*	**1A/*1G*	**3/*3*	**3/*3*	**1/*1*	**1/*1*

Table S4.4 – continued

Coriell #	*CYP2C19*		*CYP2E1*		*CYP3A4*		*CYP3A5*		*CYP4F2*	
	GeT-RM	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**	**GeT-RM**	**StellarPGx**
NA18952	**1/*1*	**1/*38*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA18959	**1/*1*	**1/*38*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1G*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA18966	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1 (*16)*	**16/*1*	**1/*3*	**1/*3*	**1/*3*	**1/*3*
NA18973	**1/*1*	**1/*1*	**1 / *7*	**1A/*7C*	**1/*1*	**1A/*1A*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA18980	**1/*1*	**1/*1*	*(*7 / *7)*	PNA	**1/*1*	**1A/*1G*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA18992	**1/*1*	**1/*1*	*(*7) / *7*	PNA	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA19003	**1/*2*	**1/*2*	**1 / *7*	**1B/*7C*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA19007	**1/*1*	**1/*1*	*(*5)/ *7*	**5B/*7A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*3*	**1/*3*
NA19095	**1/*1*	**1/*1*	**1 / *7*	PNA	**1/*1B*	**1B/*1*	**1/*3*	**1/*3*	**1/*3*	**1/*3*
NA19109	**17/*17*	**17/*17*	**1 / *1*	**1B/*1B*	**1B/*1B (*15)*	**15B/*1B*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA19122	**1 (*15)/*2*	PNA	**7 / *7*	PNA	**1B/*1B*	**1B/*1B*	**1/*1*	**1/*1*	**1/(*2)*	**1/*2*
NA19143	**1/*15*	**1/*15*	*(*7 / *7)*	PNA	**1B/*1B*	**1B/*1B*	**6/*7*	**6/*7*	**1/*1*	**1/*1*
NA19147	**1/*17*	**1/*17*	*(*4) / *7*	PNA	**1/*1B*	**1B/*1*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA19174	**1/*2*	**1/*2*	*(*7 / *7)*	PNA	**1B/*1B*	**1B/*1B*	**1/*6*	**1/*6*	**1/*1*	**1/*1*
NA19176	**2/*17*	**17/*2*	*(*7) / *7*	PNA	**1B/*1B*	**1B/*1B*	**1/*3*	**1/*3*	**1/*1*	**1/*1*
NA19178	**1 (*27)/*6*	**1/*6*	*(*7) / *7*	PNA	**1B/*1B*	**1B/*1B*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA19207	**2/*17*	**17/*2*	**7 / *7*	PNA	**1B/*1B*	**1B/*1B*	**3/*7*	**3/*7*	**1/*1*	**1/*1*
NA19213	**1/*15*	**1/*15*	*(*7) / *7*	PNA	**1B/*1B*	**1B/*1B*	**1/*6*	**1/*6*	**1/*1*	**1/*1*
NA19226	**1/*2*	**1/*2*	**1 / *7*	PNA	**1B/*1B (*15)*	**15B/*1B*	**1/*6*	**1/*6*	**1/(*2)*	**1/*2*
NA19238	**1/*1*	**1/*1*	*NA*	PNA	**1/*1*	**1B/*1B*	**1/*1*	**1/*1*	*NA*	**1/*2*
NA19239	**13/*17*	**13/*17*	*(*7) / *7*	PNA	**1/*1B*	**1B/*1G*	**1/*1*	**1/*1*	**1/*1*	**1/*1*
NA19789	**1/*1*	**1/*1*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1G*	**3/*3*	**3/*3*	*(*2)/*3*	**2/*3*
NA19819	**1/*17*	**1/*17*	**1 / *7*	**1B/*7A*	**1/*1B*	**1A/*1B*	**3/*6*	**3/*6*	*(*2)/*3*	**2/*3*
NA19908	**1/*17*	**1/*17*	**7 / *7*	PNA (gene_dup)	**1B/*1B (*15)*	**1B/*15B*	**1/*3*	**1/*3*	*(*2)/*3*	**2/*3*
NA19917	**1 (*15; *28)/*2*	**1/*2*	*NA*	PNA	**1/*1*	**1B/*1G*	**1/*7*	**1/*7*	*NA*	**1/*2*
NA19920	**1/*1*	**1/*1*	*(*4) / *7*	PNA	**1B/*1B*	**1B/*1B*	**7/*7*	**7/*7*	**1/*1*	**1/*1*
NA20296	**1/*1*	**1/*1*	**1 / *1*	**1A/*1B*	**1B/*1B*	**1B/*1B*	**1/*6*	**1/*6*	**1/*1*	**1/*1*
NA20509	**2/*2*	**2/*2*	**1 / *1*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	**1/*1*	**1/*1*
NA21781	**1/*2*	**1/*2*	*NA*	**1A/*1A*	**1/*1*	**1A/*1A*	**3/*3*	**3/*3*	*NA*	**1/*1*
NA12891	*NA*	**2/*2*	*NA*	**1A/*1A*	*NA*	**1A/*1A*	*NA*	**3/*3*	*NA*	**1/*1*
NA19240	*(*1/*17)*	**1/*17*	*NA*	PNA	**1/*1*	**1B/*1B*	**1/*1*	**1/*1*	*NA*	**1/*1*

PNA, Possible novel alleles detected; NA, Truth call not publicly available; (*n/*n or *m/*m) — Ambiguous call; *n/*n — Discordant call;
. [*m/*m] — Discordant cases where StellarPGx identified allele defining variants that were either not identified or not tested by the genotyping panels used by GeT-RM

Notes:
- The GeT-RM calls included here correspond to the reference materials characterised by Pratt et al. (2016) Notably, some alleles have now been redefined by PharmVar as outlined below.
- *CYP2C19*38* was formerly *CYP2C19*1A* (see `https://www.pharmvar.org/gene/CYP2C19`; last accessed 14-Dec-2020);
so we do not consider cases where StellarPGx called CYP2C19*38 and GeT-RM called *CYP2C19*1* as discrepancies.
- *CYP2C19*27* has been redefined as *CYP2C19*1.006* by PharmVar (see `https://www.pharmvar.org/gene/CYP2C19`; last accessed 14-Dec-2020).

Chapter 5

Characterisation of *CYP2D6* pharmacogenomic variation in African populations: An integrative bioinformatics approach

Manuscript *in preparation* by David Twesigomwe, Zané Lombard and Scott Hazelhurst. See subsection 5.7 for author contributions.

5.1 Summary

Background: *CYP2D6* is a key pharmacogene owing to its involvement in the metabolism or elimination of 25% of clinically prescribed medications and high polymorphism. Many pharmacogenetics studies have been performed worldwide to investigate the distribution of *CYP2D6* star alleles; however, African populations have been relatively understudied to date. **Methods**: In this study, the distribution of known and novel *CYP2D6* star alleles was examined across multiple sub-Saharan African populations based on the bioinformatics analysis of 961 high-depth genomes. This was followed by characterisation of novel alleles and suballeles identified in a subset of the participants via targeted long-read resequencing. **Results**: This study revealed varying frequencies of known *CYP2D6* alleles and predicted phenotypes across different African ethnolinguistic groups. In addition, 27 novel African-specific *CYP2D6* alleles were predicted computationally. Two of these novel major alleles and several suballeles were further characterised using targeted high fidelity SMRT sequencing. **Conclusion**: This study highlights the importance of studying variation in key pharmacogenes, such as *CYP2D6*, in the African context to enhance the genotyping capacity of existing panels and to support effective implementation of precision medicine in Africa. Our findings emphasise the fact that one African population should not be used as a general proxy for another — and certainly not for the whole of Africa — while developing personalised medicine strategies. This study also exemplifies the utility of high coverage whole genome sequence data and validated bioinformatics algorithms in catalysing the investigation of variations in the complex *CYP2D6* gene locus across diverse populations.

5.2 Introduction

Genetic variation is in part responsible for the inter-individual differences in drug response (Whirl-Carrillo et al., 2021). Therefore, it is important to consider an individual's genetic profile in drug therapy (along with age, sex, environment, lifestyle and other factors) in order to adjust treatment accordingly (Scott, 2011; Whirl-Carrillo et al., 2021; Zanger and Schwab, 2013). This personalised medicine approach is aimed at promoting drug efficacy and minimising the risk of adverse drug reactions. Individualising treatment relies heavily on the ability to translate information about actionable variants, haplotypes and/or diplotypes into phenotype predictions, especially for key pharmacogenes involved in the absorption, distribution, metabolism, and excretion (ADME) of drugs. However, the distribution of known alleles in such genes varies considerably within and between populations (Gaedigk et al., 2017; Y. Zhou et al., 2017). In addition, the full catalogue of pharmacogenetic variants is yet to be ascertained, even for major pharmacogenes such as *CYP2D6* (Gaedigk et al., 2021). This presents major challenges towards the application of current ADME genotyping panels and effective use of the activity score (AS) system for genotype-phenotype translation (Caudle et al., 2020; Gaedigk et al., 2008), especially in understudied populations.

CYP2D6 is a major ADME enzyme that has been widely studied given its involvement in the metabolism of 25% of clinically prescribed medications including opioids, antipsychotics, antidepressants and anticancer agents (`https://www.pharmgkb.org`). The highly polymorphic *CYP2D6* gene is located on chromosome 22q13 alongside two homologous pseudogenes, *CYP2D7* and *CYP2D8* (Kimura et al., 1989). The Pharmacogene Variation Consortium (PharmVar) has catalogued over 140 *CYP2D6* haplotypes (herein referred to as star alleles) to date (`https://www.pharmvar.org`). The functional impact of these star alleles ranges from no-function to increased function as assigned by the Clinical Pharmacogenetics Implementation Consortium (CPIC) based on published experimental studies (`https://cpicpgx.org`). *CYP2D6* genotyping is complicated due to the occurence of numerous combinations of allele-defining single nucleotide polymorphisms (SNPs) and indels (herein collectively referred to as SNVs i.e. small nucleotide variations) as well as structural variations (Nofziger et al., 2020). The *CYP2D6* copy number variations i.e. the gene deletion (*CYP2D6*5*) and gene duplication/multiplication alleles are the more common structural variations (`https://www.pharmvar.org/gene/CYP2D6`). However, other structural variants such as hybrid gene copies comprising portions from both *CYP2D6* and *CYP2D7*, often found in tandem rearrangements (duplications with non-identical gene copies) are also known to affect the CYP2D6 function, and they are difficult to investigate on available genotyping platforms (Gaedigk, 2013; Yang et al., 2017).

At present, very few African ethnolinguistic groups have been represented in *CYP2D6* pharmacogenomic studies (Gaedigk et al., 2017; see also `https://www.pharmgkb.org/page/`

cyp2d6RefMaterials). This has been in part due to lack of sufficient clinical-grade genomic datasets, and also because of the genotyping challenges associated with the complex *CYP2D6* genomic locus (Nofziger and Paulmichl, 2018). The most common African-specific *CYP2D6* star alleles include the decreased-function *CYP2D6*17* (Masimirembwa et al., 1996) and *CYP2D6 *29* (Wennerholm et al., 2001) alleles. However, recent studies on more sub-Saharan African cohorts have demonstrated unique *CYP2D6* star allele distributions and found novel, albeit relatively rare alleles. These rarer African-specific alleles include *CYP2D6*70* (Matimba et al., 2009), *CYP2D6*73* and *74* (Wright et al., 2010), *CYP2D6*84* (Dodgen et al., 2013), and *CYP2D6*106* (Dodgen et al., 2015). Some of these novel alleles e.g. *CYP2D6*73* (unknown function), *74* (unknown function) and *106* (uncertain function) represent new combinations of known function-altering variants on other existing haplotypes (https://www.pharmvar.org/gene/CYP2D6). Conversely, other alleles such as *CYP2D6*84* (Dodgen et al., 2013) were completely novel on discovery and submission to PharmVar. This underscores the need for investigation of the *CYP2D6* allelic landscape across diverse African populations in order to reliably inform the design of comprehensive genotyping panels and the much-needed clinical pharmacogenomics implementation strategies across Africa.

This study therefore aims to investigate the distribution of known and potential novel *CYP2D6* alleles across diverse African populations, and to predict the metaboliser phenotype distribution. In our analysis, we leverage the capabilities of recently developed and benchmarked star allele calling bioinformatics algorithms for determining *CYP2D6* diplotypes from short-read high coverage whole genome sequence (WGS) data. This is followed by targeted single molecule real-time sequencing (SMRT)-based characterisation of the novel alleles and suballeles present in some of the participants based on availability and access to sample DNA. A few long-read-based allele characterisation protocols have been validated for *CYP2D6* (Buermans et al., 2017; Ammar et al., 2015; Liau et al., 2019; Qiao et al., 2016). SMRT sequencing especially using the high fidelity (HiFi) read data provides the opportunity to comprehensively characterise novel *CYP2D6* alleles in African populations, which are known to have high genetic diversity compared to other global populations.

In addition to the aforementioned analyses, we compare the *CYP2D6* allele frequencies in sub-Saharan Africa to the distribution across global populations represented in the 1000 Human Genomes Project high-coverage dataset (Byrska-Bishop et al., 2021).

5.3 Materials and Methods

5.3.1 Ethics statement

This study was approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (Wits HREC-Medical) (M200993 — see Appendix B on page 215). We performed secondary analysis of data that had been generated and studied for other purposes. This included data from public repositories and data from contributing Africa-based studies/centres, each of which obtained additional local ethics approval (see subsection 5.7 on page 145).

5.3.2 Study population and whole genome sequence data sources

Table 5.1 on page 116 and Figure 5.1 provide a summary of the African whole genome sequence datasets used in this study. These data were all sequenced on Illumina platforms to a minimum depth of 30×. The bulk of the datasets are from the H3A and the Human 1000 Genomes consortia — and they mainly represent Niger-Congo ethnolinguistic groups, while there was limited representation from Nilo-Saharan and Afro-Asiatic populations and few individuals of Khoe and San heritage. The population structure for the participants in this study correlates strongly to geographical location as detailed by Choudhury et al. (2020) and da Rocha, Othman, Botha, et al. (2021).

5.3.2.1 The H3Africa-Baylor data

The H3Africa-Baylor dataset comprises 272 samples from individuals that were sequenced under various studies by the Human Heredity and Health in Africa (H3A) Consortium. These individuals are representative of populations from Benin (FNB, the Fon in Benin; n=50), Botswana (BOT; n=47), Cameroon (CAM; n=26), Nigeria (BRN, the Berom of Nigeria; n=49), Zambia (BSZ, Bantu speakers in Zambia; n=41), Ghana (GHA; n=26) and Burkina Faso (BFA; n=33). The details about the sample preparation and sequencing have been previously described (Choudhury et al., 2020; da Rocha, Othman, Botha, et al., 2021).

5.3.2.2 H3Africa AWI-Gen data

The AWI-Gen set consists of 100 south eastern Bantu-Speakers (SEB) from South Africa. AWI-Gen stands for the the Africa Wits-INDEPTH partnership for Genomic Studies (Ramsay et al., 2016). DNA samples from the 100 SEB AWI-Gen participants, and aliquots of 59 samples provided by AWI-Gen to H3A Baylor (GHA and BFA participants) were available at the Sydney Brenner Institute for Molecular Bioscience; therefore, they were included for the laboratory validation of *CYP2D6* alleles (subsubsection 5.3.5) under ethics amendment terms in protocol M200993.

Figure 5.1: Geographical regions and African populations represented in the study. The size of the circles reflect the relative number of participants from each population group. AWI-Gen, Africa Wits-INDEPTH partnership for Genomics studies; CBRL, Cell Biology Research Laboratory (National Institute of Communicable Diseases, Wits); H3A, Human Heredity and Health in Africa Consortium; SAHGP, Southern African Human Genome Programme.

Table 5.1: Sources of the high coverage WGS datasets used in the study. The genomes of all the study participants were sequenced to an average read depth of ∼30x by the respective projects indicated in the table.

Populations	Project/Institution	*n*
H3Africa Consortium data		
The Fon from Benin (FNB)	H3A Baylor; University of Montréal	50
Berom of Nigeria (BRN)	H3A Baylor; Institute of Human Virology	49
Cameroon (CAM)	H3A Baylor; University of Dschang	26
Ghana (GHA)	H3A Baylor; AWI-Gen	26
Burkina Faso (BFA)	H3A Baylor; AWI-Gen	33
South Africa	AWI-Gen	100
Botswana (BOT)	H3A Baylor; CAfGEN	47
Bantu-speakers from Zambia (BSZ)	H3A Baylor; University of Zambia	41
Data from other Africa-based projects		
South Africa	SAHGP	15
South Africa	CBRL (NICD, Wits)	40
Public datasets with African genomes		
Algeria	SGDP [a]	1
Botswana/Namibia	SGDP	3
Namibia	SGDP [b]	3
DRC	SGDP [c]	4
Gambia	SGDP	2
Kenya	SGDP [d]	5
Nigeria	SGDP	4
Senegal	SGDP	3
South Africa	SGDP [b]	3
South Sudan	SGDP [d]	3
Luhya in Webuye, Kenya (LWK)	1000 Genomes Project	99
Esan in Nigeria (ESN)	1000 Genomes Project	99
Yoruba in Ibadan (YRI)	1000 Genomes Project	108
Mende in Sierra Leone (MSL)	1000 Genomes Project	85
Gambian in Western Division, Mandinka (GWD)	1000 Genomes Project	113
Public datasets with genomes from other global superpopulations (for comparative analysis)		
African Caribbean in Barbados (ACB)	1000 Genomes Project	96
People with African Ancestry in Southwest USA (ASW)	1000 Genomes Project	61
European (EUR)	1000 Genomes Project	503
Admixed American (AMR)	1000 Genomes Project	347
South Asian (SAS)	1000 Genomes Project	489
East Asian (EAS)	1000 Genomes Project	504

AWI-Gen, Africa Wits-INDEPTH partnership for Genomics studies; CAfGen, The Collaborative African Genomics Network; CBRL, Cell Biology Research Laboratory; NICD, National Institute for Communicable Diseases; SAHGP, Southern African Human Genome Programme; SGDP, Simons Genome Diversity Project. The majority of the continental African participants referred to in this table are from the Niger-Congo language family, except the following: [a] Afro-Asiatic speaker; [b] Khoe and San Hunter-gatherers; [c] Pygmy Hunter-gatherers; [d] Nilo-Saharan (n=2 for the Kenyan SGDP participants)

5.3.2.3 Additional data from projects/labs led by African collaborators

The dataset from the Cell Biology Research Laboratory (CBRL) consists of 40 samples from South African participants (39 black and one of mixed ancestry). The details about the sample preparation and Illumina sequencing have been previously described by da Rocha, Othman, Botha, et al. (2021). Through the collaboration with CBRL, DNA samples from these 40 participants were included in the long-read-based *CYP2D6* allele validation.

The SAHGP set comprises 24 WGS samples (at least 30× read depth, 101 bp paired-end reads and ∼314 bp insert size) generated by the Southern African Human Genome Programme (Choudhury et al., 2017). Fifteen SAHGP participants included in this study are from two Bantu-speaking ethnolinguistic groups: Sotho-Tswana speakers (n = 8) and the Xhosa speakers (n = 7) (Choudhury et al., 2017). We excluded participants (n=7) of recent admixed ancestry.

5.3.2.4 Publicly available African WGS data

We included 31 high-coverage (average depth of 43×) African genomes generated by the Simon's Genome Diversity Project (SGDP) (Mallick et al., 2016) in the study. These samples are from individuals representative of various ethnolinguistic groups in Congo, Namibia, Kenya, Senegal, Algeria, Nigeria, Gambia, Sudan and South Africa.

The 1000G_AFR set consists of 504 African participants from the continental African populations represented in the 1000 Genomes Project. These include Gambian Mandinka (GWD; n=113), Mende from Sierra Leone (MSL; n=85), Yoruba from Ibadan, Nigeria (YRI; n=108), the Esan from Nigeria (ESN; n=99) and the Luhya from Webuye, Kenya (LWK; n=99). The samples from these individuals have recently been sequenced to a mean read-depth of 34×. WGS libraries were prepared using the Truseq DNA PCR-free (450bp) Library Preparation Kit and then sequenced (2 x 150bp cycles) on an Illumina Novaseq 6000 sequencer after additional processing.

We also included the rest of the 1000 Genomes Project samples in our analysis for comparison of the *CYP2D6* allele distribution in Africa versus that in global populations. These data are representative of European (n=503), East Asian (n=504), South Asian (n=489), admixed American (n=347),

5.3.2.5 Other Metadata

For the participants included in this study, the only phenotype disclosed to us by the primary research projects was sex. The self-identified ethnicity, location in the country, and disease status were not revealed.

5.3.3 Star allele analysis

Given the known *CYP2D6* genotyping complexity, we called *CYP2D6* star alleles from the African WGS samples using five separate tools including StellarPGx (v1.2.4), Cyrius (v1.1) (Chen et al., 2021), Aldy (v3.1) (Numanagić et al., 2018), Stargazer (v1.0.8) (S.-B. Lee, Wheeler, Thummel, et al., 2019) and Astrolabe (v0.8.7.0) (Twist et al., 2016). Thereafter, we obtained a consensus *CYP2D6* diplotype call-set from the output of these algorithms (Figure 5.2). In a few cases e.g. samples for which no concordant diplotype was obtained between at least 2 of the tools and samples that seemed to have challenging structural rearrangements between *CYP2D6* and *CYP2D7*, we performed manual visual inspection of the read depth plots output by Stargazer and SAMtools depth (H. Li and Durbin, 2009).

StellarPGx, Cyrius, and Aldy required only the Binary Alignment (BAM) file as input and called sample diplotypes from data aligned to both the GRCh37 and GRCh38 reference genomes. For Astrolabe and Stargazer, the required Variant Call Format (VCF) files were generated using the H3ABioNet's GATK pipeline (`https://github.com/h3abionet/recalling`). Briefly gVCFs were called using HaplotypeCaller and then combined using CombineGVCF in GATK v4.0.8.1 (Van der Auwera et al., 2013). Raw variants were joint-called using GenotypeGVCFs (v4.1.3.0). The SNPs and indels in the WGS VCFs were subjected to Variant Quality Score Recalibration, and thereafter the high quality (PASS) variants corresponding to the *CYP2D6* loci were extracted from the WGS VCF using BCFtools. Astrolabe and Stargazer also required the BAM file and 'GATK-DepthOfCoverage Format' (GDF) files, respectively, for depth of coverage analysis and structural variant calling. The *VDR* gene was used as the control locus for Stargazer's star allele analysis.

Stargazer v1.0.8 does not have support for GRCh38. The GRCh38-aligned WGS data included in this study was the 1000 Genomes WGS data. In order to run Stargazer on this data, we extracted reads aligned to the *CYP2D*, and *VDR* (control gene locus) GRCh38 regions using SAMtools, and realigned them to GRCh37 using BWA-MEM. The variant calling and depth of coverage analysis was similar to the aforementioned steps for the WGS data except that we performed hard-filtering in the variant quality control for the realigned regional data. Stargazer was then run on the *CYP2D6* VCF files and the required GDF files for these samples.

5.3.4 Metaboliser phenotype prediction

The *CYP2D6* consensus diplotype calls were translated to phenotypes by using the standardised activity score system recommended by CPIC (Caudle et al., 2020). For each haplotype, activity values were assigned as follows: 0, 0.5, 1 and >1 (dependent on gene copies) for no-function, decreased function, normal function and increased function haplotypes, respectively.

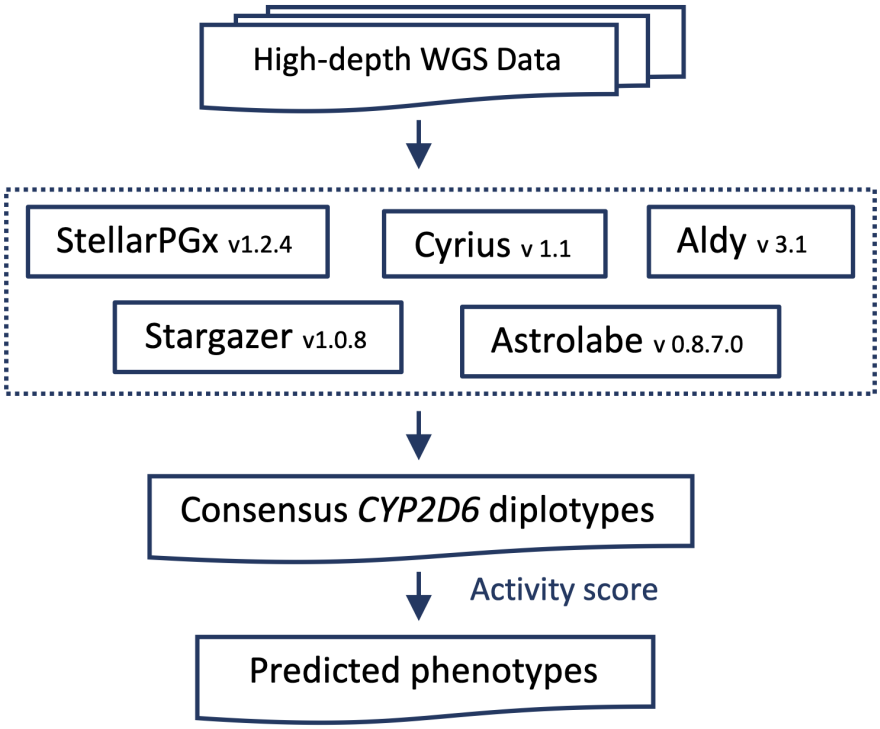


Figure 5.2: *CYP2D6* Star allele calling workflow overview

The *CYP2D6*10* allele was assigned an activity value of 0.25 (Caudle et al., 2020). The activity score for each diplotype was obtained by adding up the activity values of component haplotypes. The participants were then classified as either poor metabolisers (AS=0), intermediate metabolisers ($0<AS<1.25$), normal metabolisers ($1.25\leq AS\leq 2.25$) or ultrarapid metabolisers ($AS>2.25$) (Caudle et al., 2020).

5.3.5 Targeted long-read resequencing

5.3.5.1 *CYP2D6* long-range PCR

CYP2D6 6.6kb-long amplicons (FragA, Figure 5.3) that contain the *CYP2D6* gene as well as upstream and downstream non-coding regions were generated following the long-range PCR (XL-PCR) protocol described by Gaedigk et al. (2007) with some modifications (Table 5.2 and Table 5.3). The XL-PCR forward and reverse primers were tailed with universal sequences (5'-GCAGTCGAACATGTAGCTGACTCAGGTCAC-3' and 5'-TGGATCACTTGTGCAAGCAT-CACATCGTAG-3', respectively) on the 5' end to enable sample barcoding of pooled amplicons via a second PCR and subsequent multiplexing. For samples where a gene duplication was computationally-inferred from the short read WGS, the duplicated *CYP2D6* gene copy was amplified with FragD primers (Gaedigk et al., 2007) which produced 8.6 kb fragments (Figure 5.3).

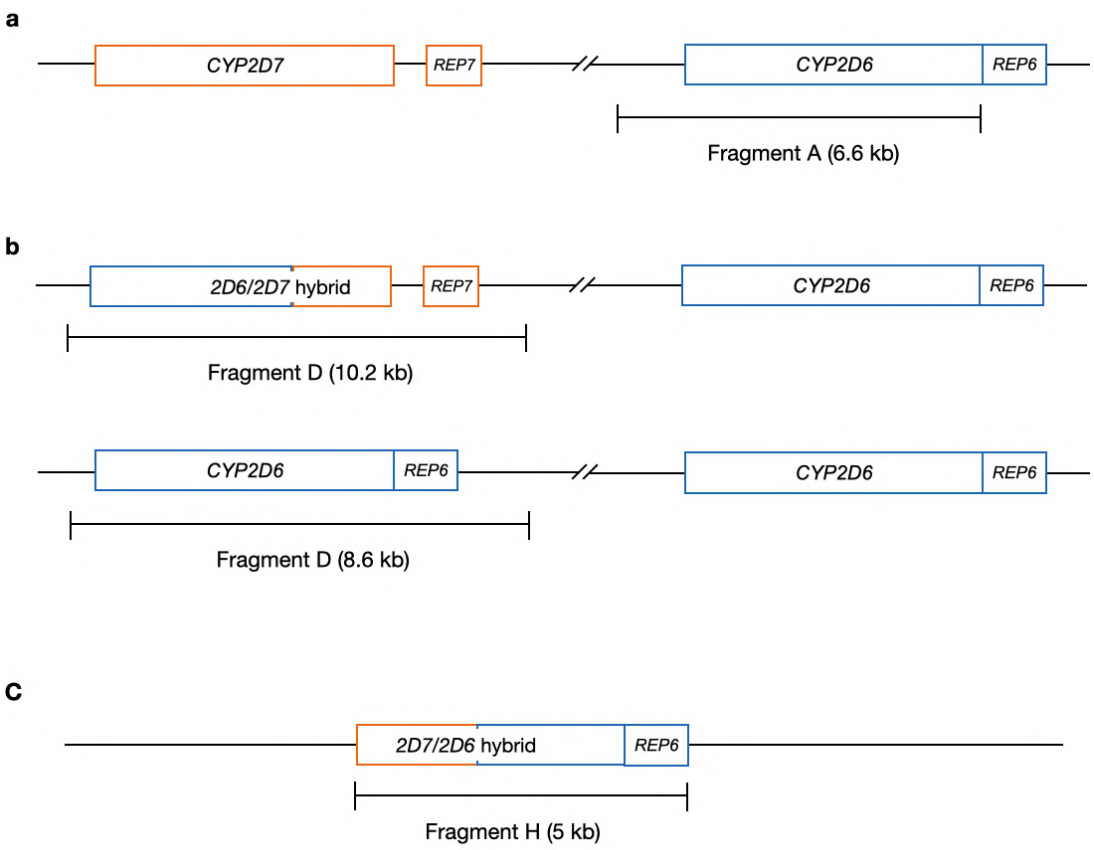


Figure 5.3: *CYP2D6* XL-PCR fragments generated for SMRT sequencing. (**a**) Depicts the arrangement without structural variations (except for the intron 1 conversion in some cases). (**b**) Depicts Fragment D (FragD) which may arise due to duplication of a normal *CYP2D6* copy (8.6 kb fragment) or due to *CYP2D6-2D7* hybrid formation (10.2 kb fragment). (**c**) Graphical representation of a *CYP2D7-2D6* hybrid fragment (*CYP2D6*13* alleles).

Table 5.2: Primers used for generating *CYP2D6* long-range PCR fragments in this study

Assay	Primer name	Direction	5'–sequence	References
CYP2D6 FragA (6.6 kb)	5'2D6 6F	F	TCACCCCCAGCGGACTTATCAACC	(Gaedigk et al., 2007)
	3'2D6 R/3	R	CGACTGAGCCCTGGGAGGTAGGTAG	(Gaedigk et al., 2007)
CYP2D6 FragD (8.6 kb or 10.2 kb)	5'2D6 F/2	F	CCAGAAGGCTTTGCAGGCTTCAG	(Gaedigk et al., 2007)
	3'2D6x2 R/3	R	CGGCAGTGGTCAGCTAATGAC	(Gaedigk et al., 2007)
CYP2D6 FragH (∼5 kb)	5'2D7 XL-5F	F	TCCGACCAGGCCTTTCTACCAC	(Gaedigk and Coetsee, 2008)
	3'2D6 R/3	R	CGACTGAGCCCTGGGAGGTAGGTAG	(Gaedigk and Coetsee, 2008)

F, Forward; R, Reverse; ex, exon; int, intron

FragD primers are also used to amplify *CYP2D6-2D7* hybrid alleles (∼10.2kb). For samples where *CYP2D6*13* was predicted, FragH primers (Gaedigk et al., 2007) were used to amplify the ∼5kb-long *CYP2D7/2D6* hybrid amplicons. Each 20 µl reaction mix contained 60-120 ng of genomic DNA, 10.0 µl of 2X LongAmp Taq ReadyMix (New England Biolabs, South Africa), 1 µl of 100% DMSO (Sigma-Aldrich/Merck, South Africa), and 1.0 µl each of 10 µM forward and reverse primers (Inqaba Biotech, South Africa). The details about the specific primers used to amplify the aforementioned XL-PCR fragments and the PCR cycling conditions are provided in Table 5.2 and Table 5.3.

Table 5.3: PCR cycling conditions for generating *CYP2D6* fragments (first amplification)

	CYP2D6 assays		
	1	2	3
First denaturation (ºC:s)		94:180	
Denaturation (ºC:s)	94:20	94:20	94:20
Annealing (ºC:s)	68:30	68:30	68:30
Extension (ºC:s)	68:420	68:780	68:420
Extension (ºC:s)	68:420	68:780	68:420
Final extension (ºC:s)	68:420	68:780	68:420
Hold		10ºC	

[1] Amplification for *CYP2D6* FragA.
[2] Amplification for *CYP2D6* FragD.
[3] Amplification for *CYP2D6* FragH.

5.3.5.2 Amplicon pooling and barcoding

Quality control of the PCR products from the first-round PCR was carried out using a 0.8% agarose gel for visual inspections and the Agilent 4200 TapeStation (Diagnostech, South Africa) for quantification following the D5000 Screen tape kit. For each participant, *CYP2D6* amplicons were pooled equimolarly with *CYP2A6* and *CYP2B6* amplicons from a related subproject (Chapter 6). All amplicons were in the size range of 5–12 kb. The amplicon pools were purified using the AMPure PB bead purification (`https://www.pacb.com`). Thereafter, barcodes were added to the purified amplicons via a second-round PCR. The 25 µl reaction mix contained 5-10 ng/µl of first-round pooled PCR product, 11.0 µl of 2X longAmp Taq ReadyMix (New England Biolabs, South Africa), 1 µl of DMSO (100%), and 5.0 µl of 2 µM barcoded universal primers. The PCR cycling conditions were as follows: 20 cycles of 95°C for 1 minute, 65°C for 30 seconds, 72°C for 11 minutes.

5.3.5.3 PacBio sequencing and data analysis

As with amplicon pooling and barcoding, HiFi sequencing (Pacific Biosciences) was outsourced to Inqaba Biotech (Pretoria, South Africa). SMRTBell libraries were constructed from ∼500 ng of pooled fragments by following standard end-repair, adapter ligation, and purification strategies detailed in the Pacific Biosciences protocols (`https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf`). Annealing and binding of SMRTbell templates was performed using the Sequel II Binding kit 2.2 and sequencing primer v5, and Circular Consensus Sequencing (CCS) was performed for a movie time of 30hrs to generate High Fidelity (HiFi) reads via the SMRT Link software on the Sequel IIe instrument (`https://www.pacb.com`) at Inqaba Biotech (Pretoria, South Africa).

Raw HiFi sequence data were demultiplexed and processed into individual samples according to the corresponding barcode sequences using the NGSutils next-generation sequencing data analysis software kit (Breese and Liu, 2013). *CYP2D6* HiFi reads were aligned to the *CYP2D6* region in GRCh38 and GRCh37, and also to the NG_008376.4 (LRG_303) reference sequence using pbmm2 v1.7.0 (`https://github.com/PacificBiosciences/pbmm2`). Variant calling was done using DeepVariant (Poplin et al., 2018). Thereafter, variant phasing and read haplotagging was carried out for samples containing more than one heterozygous variant using WhatsHap (Patterson et al., 2015).

5.3.6 Variant functional prediction

CYP2D6 variants were annotated using the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016). The functional effects of potential novel star allele-defining variants were predicted using VEP plugins including SIFT (Kumar et al., 2009), Polyphen-2 (Adzhubei, Jordan, and Sunyaev, 2013), CADD (Rentzsch et al., 2019), LRT (Chun and Fay, 2009), PROVEAN (Choi et al., 2012), and VEST4 (Carter et al., 2013), taking into account the ADME-optimised parameters suggested by Y. Zhou et al. (2019). SIFT Indel (Hu and Ng, 2013) was used to annotate frameshift variants.

5.3.7 Statistical analysis

Allele frequencies were calculated and tested for deviation from Hardy Weinberg equilibrium (HWE) using the Fisher's exact test in R (`https://www.r-project.org`). A Pearson χ^2 analysis was used to determine significant differences in population *CYP2D6* allele and diplotype frequencies. P-values of < 0.05 were considered statistically significant. Allele frequency data for previous studies across SSA was obtained from the Pharmacogenomics Knowledgebase (PharmGKB) (`https://www.pharmgkb.org/page/cyp2d6RefMaterials`).

5.4 Results

5.4.1 Sample populations

Our study dataset was mainly representative of Niger-Congo ethnolinguistic groups (Table 5.1) (see also Choudhury et al., 2020; da Rocha, Othman, Botha, et al., 2021). Only a few genomes from Nilo-Saharan (3) and Afro-Asiatic (1) populations were analysed in this study. In particular, the SGDP dataset included a few individuals of Khoe and San heritage. Significant admixture from Khoe and San speakers is known to occur among Bantu-speakers in Southern Africa (Sengupta et al., 2021). Results from the *CYP2D6* star allele analysis of African American/Afro-Caribbean (ASW and ACB), European (EUR), admixed American (AMR), South Asian (SAS), and East Asian (EAS) populations represented in the 1000 Genomes Project high coverage dataset (Byrska-Bishop et al., 2021) are included for comparison.

5.4.2 *CYP2D6* star allele frequencies

For previously characterised alleles, the nomenclature used here is in accordance with PharmVar definitions (`https://www.pharmvar.org/gene/CYP2D6`). All *CYP2D6* variant denotation follows the numbering based on NG_008376.4 counting from the sequence start, and/or protein change where applicable.

Of the known *CYP2D6* star alleles catalogued by PharmVar (`https://www.pharmvar.org/gene/CYP2D6`), 38 distinct alleles were detected in the sub-Saharan African datasets in this study, including 11 *CYP2D6* structural variants (SVs) (Table 5.4 and Table 5.5). These *CYP2D6* SVs (i.e. **1xN*, **2xN*, **4xN*, **5*, **13*, **17x2*, **29x2*, **36*, **43x2*, **45xN*, **68+*4*) accounted for 14% of the *CYP2D6* haplotypes called from SSA populations (Figure 5.4). Notably, some alleles considered under the SNV-defined allele bracket are known to harbour smaller benign SVs e.g. the *CYP2D6* intron 1 conversion to *CYP2D7* (`https://www.pharmvar.org/gene/CYP2D6`). Of the resident African study participants, 27% harboured at least one SV-defined *CYP2D6* allele. This was comparable to the proportion of individuals with *CYP2D6* SVs in the ASW (29.5%) and ACB (28%) populations, higher than the proportion in European participants (22.7%), significantly higher than the proportion in admixed Americans (17%, p-value = 0.00015) and South Asians (14.1%, p-value = 1.4e-08), and significantly less than the proportion in East Asians (66.3%, p-value $<$ 2.2e-16).

The Algerian participant (LP6005441-DNA_G08, SGDP dataset) had the *CYP2D6*35/*4* diplotype, and was excluded from subsequent comparative analyses as this was the only resident African participant from outside SSA in this study. The *CYP2D6*35* allele was not detected in the cohorts from within SSA.

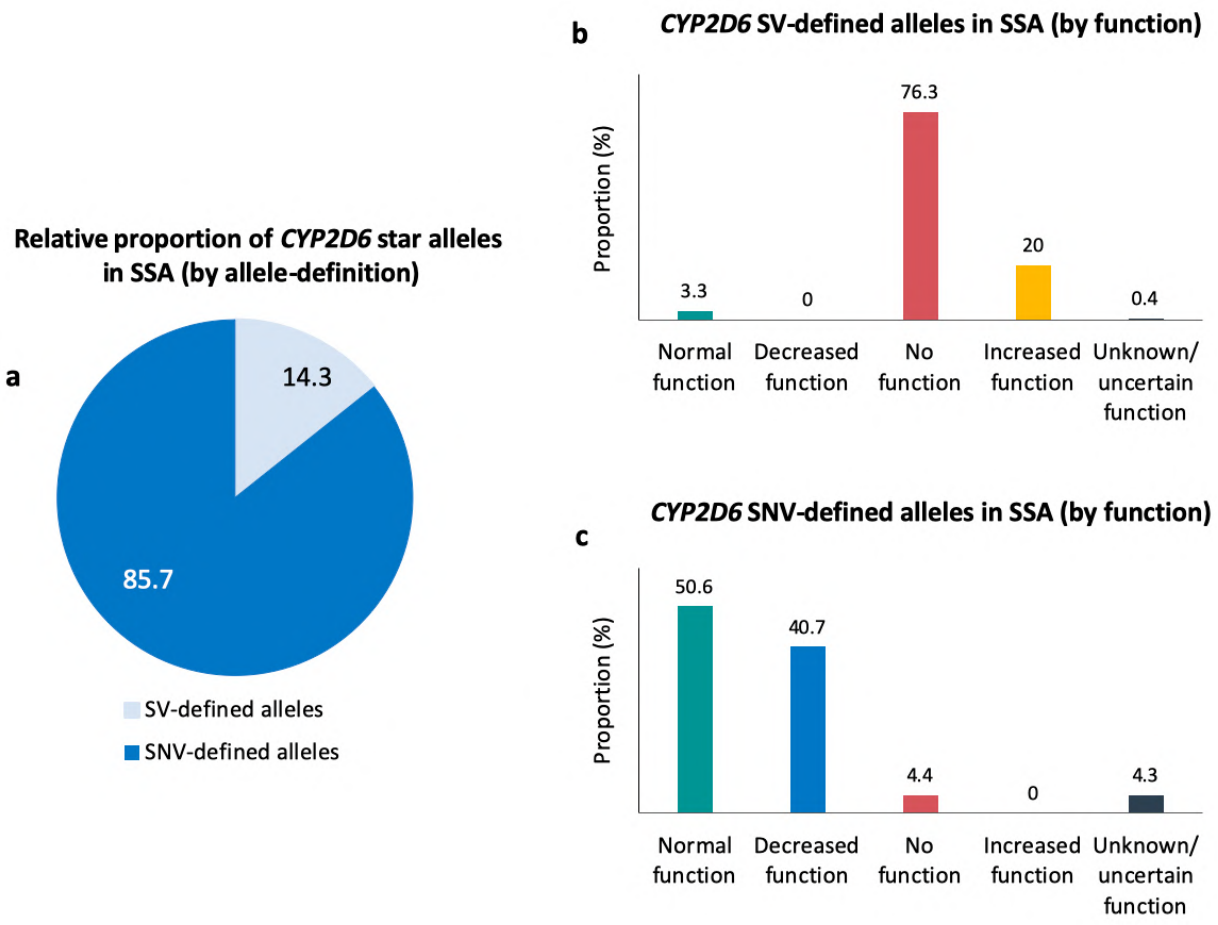


Figure 5.4: Proportion of *CYP2D6* alleles defined by structural variants (270 alleles) and core SNVs (1598 alleles) in sub-Saharan African populations. Panel (**a**) shows the relative proportion *CYP2D6* star alleles defined by structural variations (SVs) — e.g. the gene deletion (**5*), gene duplications, hybrids and tandems — compared to the relative proportion alleles defined by various combinations of key SNVs including rs28371706 (**17*), rs61736512 and rs59421388 (**29*), and rs1065852 (**10*). Panel (**b**) depicts the relative proportion of structural variants by function across SSA. *CYP2D6*5* accounts for the high proportion of no function alleles. Normal function SV-defined alleles arise from duplication of decreased function SNV-defined alleles e.g. *CYP2D6*17x2* and **29x2*, while increased function SV-defined alleles mainly include duplications of normal function alleles e.g. *CYP2D6*1xN* and **2xN*. Panel (**c**) shows the relative proportions of SNV-defined alleles grouped by CPIC function. *CYP2D6*17* and **29* account for the high proportion of decreased function alleles across SSA.

Among the known normal function alleles, *CYP2D6*1* (reference allele, but known to have many suballeles) was the most frequent in SSA followed by *CYP2D6*2* and *CYP2D6*45* (Table 5.4). The *CYP2D6*1* frequency observed in SSA in this study (24.7%) is higher than the average frequency (7.8%) from previous SSA studies curated by PharmGKB, while the reverse is true for the observed SSA *CYP2D6*2* frequency (13.4%) in this study. *CYP2D6*45* is largely African-specific as expected (Table 5.4). The *CYP2D6*17x2* and *CYP2D6*29x2* structural variant alleles found in SSA are grouped under normal function alleles despite **17* and **29* being decreased function alleles (see subsection 5.5).

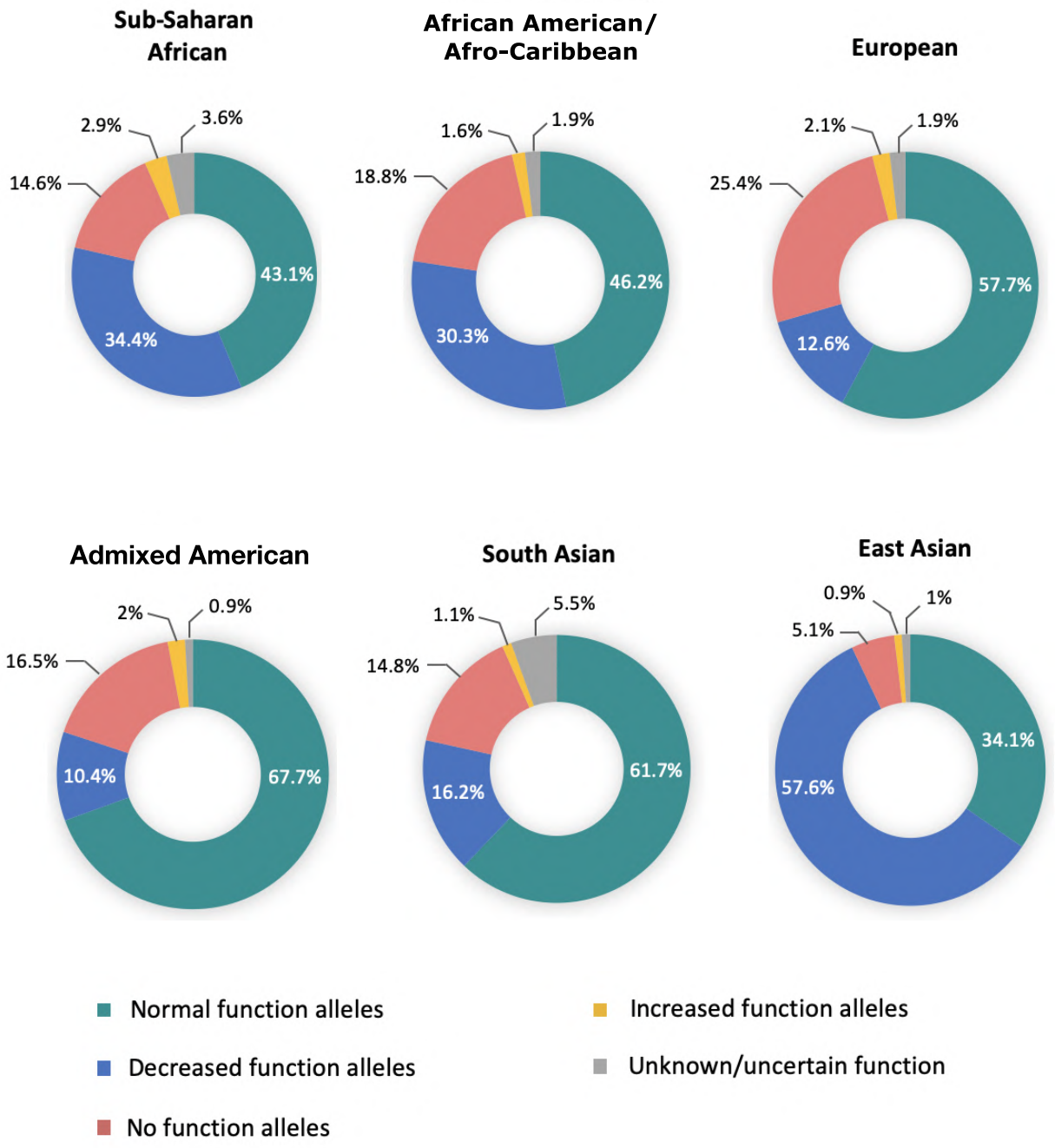


Figure 5.5: Frequency distribution of *CYP2D6* star alleles grouped by function across African and global populations. SSA populations have a higher frequency of decreased function *CYP2D6* star alleles compared to European, admixed American and South Asian populations in part due to a higher frequency of *CYP2D6*17* and **29*. Conversely, *CYP2D6*10* accounts for the high proportion of decreased function alleles in East Asian populations. SSA and South Asian populations have a relatively higher proportion of alleles with unknown or uncertain function — which warrants more functional assays to determine the CPIC function of these alleles.

We detected *CYP2D6*17*, **29*, **41* and **84* in SSA among the known decreased function *CYP2D6* alleles. As expected, *CYP2D6*17* which is defined mainly by 6041C>T (T107I) was the most frequent decreased function allele found in SSA participants and among the African Americans/Afro-Caribbeans, followed by *CYP2D6*29* (Table 5.4). *CYP2D6*17* was found to be most frequent among the Fon from Benin (AF = 30%) and significantly less frequent among the participants from Cameroon (AF = 5.8%) compared to participants from both neighbouring and distant groups within SSA (Table 5.5). *CYP2D6*29*, which is defined mainly by 6679G>A + 6681G>C (V136I), and 8203G>A (V338M) was also non-uniformly distributed across SSA (Table 5.5). In addition the frequency of *CYP2D6*29* was found to be significantly lower (p-value = 0.01598) in ACB and ASW (combined AF = 5.7%) compared to the average frequency in SSA Table 5.4. Notably, the observed frequency of *CYP2D6*41* in our SSA cohorts (AF=0.8%) was significantly less than the average frequency (11.5%) calculated by the PharmGKB from previous studies in the region (PharmGKB and CPIC *CYP2D6* reference materials, 2021). Also, none of the SSA participants with the *CYP2D6*10* allele had the **36* allele in tandem.

Among the no function *CYP2D6* alleles, we found the *CYP2D6* full gene deletion (*CYP2D6*5*) to be more frequent in SSA (AF = 8.1%) compared to previously reported frequencies (Gaedigk et al., 2017) and to all the other biogeographical groups (Table 5.4). *CYP2D6*5* accounted for over 52% of the no function haplotypes called from SSA populations, and it occured most frequently among the participants from Botswana (AF = 13.8%) and South Africa (AF = 14.7%), but it was less frequent among the East and West African population groups e.g. LWK (AF = 4.7%) and ESN (AF = 3.5%), respectively (Table 5.5). In addition, the *CYP2D6*5/*5* diplotype (i.e. homozygous *CYP2D6* gene deletion) was observed in 10 SSA individuals, including 3 from Botswana and 3 from South Africa, but it was virtually absent across the other biogeographical groups. The *CYP2D6*4* allele defined by rs3892097 (splice defect) occurred at a significantly less frequency (1.7%) in the SSA populations in this study compared to the frequency in Europeans (AF=11.8%), and was also less frequent than the estimated average frequency from previous studies in SSA (p-value = 0.0002) (Table 5.4). Conversely, we found *CYP2D6*4xN* alleles to be more frequent in SSA as compared to European populations, which have a higher frequency of *CYP2D6*68+*4* and the rarer *CYP2D6*4+*4.013* (Table 5.4). The African-specific no function alleles called in SSA diplotypes were *CYP2D6*12*, *CYP2D6*15 CYP2D6*40*, **42* and **56*. The no function hybrid alleles which were detected, albeit rarely in SSA include *CYP2D6*13*, *CYP2D6*36*, and *CYP2D6*68+*4* (Table 5.4 and Table 5.5).

The increased function *CYP2D6* star alleles identified in SSA in this study include *CYP2D6*1xN*, **2xN*, and **45xN* (Table 5.4 and Table 5.5). These accounted for 20% of the *CYP2D6* structural variants identified in SSA (Figure 5.4) and were generally more frequent in western African populations compared to southern African populations (Table 5.5). For example, *CYP2D6*2x2* was absent from Zambian and South African participants but had a frequency of $\geq$2% among

the FNB, BFA, CAM, ESN, YRI, and GWD populations. However, this allele was found to be frequent in Botswana (AF = 3.2%) as well.

Regarding the *CYP2D6* alleles with unknown/uncertain function, these were more frequent in SSA (AF = 3.6%) and South Asian (AF = 5.5%) populations compared to other biogeographical groups (Figure 5.5). The African-specific alleles in this category were the relatively common *CYP2D6*106* (SSA AF=1.5%), and also **73*, **74*, **125* and **149* which were rare (Table 5.4). The *CYP2D6*43* allele was most frequent among the LWK (AF = 3.2%) compared to other SSA populations (Table 5.5).

Overall, SSA populations were found to have a statistically significant higher frequency of decreased function *CYP2D6* alleles compared to European (p-value = 2.2e-16), admixed American (p-value = 2.2e-16) and South Asian (p-value = 2.2e-16) populations (Figure 5.5) largely due to the high prevalence of the African-specific *CYP2D6*17* and **29* alleles. However, East Asian populations had the highest frequency of decreased function *CYP2D6* alleles (57.6%) mainly due to the high occurrence of *CYP2D6*10* which is often involved in various tandem rearrangements with the **36* hybrid (Table 5.4).

Table 5.4: *CYP2D6* allele frequencies in sub-Saharan African populations compared to previous studies in this region and to the distribution in global populations.

		Allele frequencies in indicated populations (%)						
***CYP2D6* allele**	**CPIC function**	SSA in this study (n=947)	Previous SSA studies [a] (n=2248)	African American/ Afro-Caribbean (n=157)	EUR (n=498)	AMR (n=345)	SAS (n=481)	EAS (n=502)
**1*	Normal	24.7	7.8	31.5	35.8	45.1	39.5	26
**2*	Normal	13.4	19.8	9.6	15.9	18.3	20.8	7.8
**13+*1*	Normal	0		0	0.1	0	0	0
**13+*2*	Normal	0		0	0.2	0.4	0.1	0
**27*	Normal	0.3	0.5	0.3	0	0.3	0	0
**33*	Normal	0		0	0.6	0.3	0.2	0
**34*	Normal	0.1		0	0	0	0	0
**35*	Normal	0	0	1	4.8	2.6	0.6	0
**39*	Normal	0.1	0	0.3	0.1	0.1	0.4	0.3
**45*	Normal	3.7	4.2	2.2	0	0.3	0	0
**46*	Normal	0.5	0.2	1	0	0.1	0	0
**17x2*	Normal	0.2		0	0	0	0	0
**29x2*	Normal	0.3		0.3	0	0	0	0
**9x2*	Normal	0		0	0.1	0	0	0
**41x2*	Normal	0	0.2	0	0	0	0	0
**68+*2*	Normal	0		0	0.1	0	0.1	0
**83+*2*	Normal	0		0	0	0.1	0	0
**9*	Decreased	0	0	0.3	2.4	1.3	0	0
**10*	Decreased	3.9	5.6	3.8	1.2	1.6	3.5	15.2
**10x2*	Decreased	0	0	0	0	0	0	0.4
**14*	Decreased	0		0	0	0	0	1
**17*	Decreased	19.5	19.3	16.9	0.2	0.9	0	0
**29*	Decreased	10	12.1	5.7	0	0.3	0	0
**36+*10*	Decreased	0		0.3	0	0.1	1.2	34.5
**36+*10.003*	Decreased	0		0	0	0	0	0.1
**36x2+*10*	Decreased	0		0	0	0	0	1.7
**36+*10x2*	Decreased	0		0	0	0	0	0.5
**36+*10≥3*	Decreased	0		0	0	0	0	0.1
**36x2+*10x2*	Decreased	0		0	0	0	0	0.1
**36x2+*10≥3*	Decreased	0		0	0	0	0	0.1
**41*	Decreased	0.8	11.5	2.9	8.5	6.1	11.4	3.4
**49*	Decreased	0	0	0	0	0	0	0.5
**59*	Decreased	0		0	0.2	0.1	0	0
**84*	Decreased	0.1		0.3	0	0	0	0
**3*	No function	0	0.2	1	1.8	0.6	0.2	0
**4*	No function	1.7	3.4	4.1	11.8	9.6	8.2	0.2

Table 5.4 – continued

CYP2D6 allele	CPIC function	Allele frequencies in indicated populations (%)						
		SSA in this study (n=947)	Previous SSA studies [a] (n=2248)	African American/ Afro-Caribbean (n=157)	EUR (n=498)	AMR (n=345)	SAS (n=481)	EAS (n=502)
**4x2*	No function	2.2		3.8	0.3	0.1	0	0
**4x3*	No function	0.1		0.1	0	0	0	0
**4+*4.013*	No function	0		0	0.7	0	0	0
**5*	No function	8.1	5.2	7	2.4	2.2	2.5	3.5
**6*	No function	0	0	0.3	2.1	0.3	0.1	0
**7*	No function	0	0	0	0	0	0.9	0
**11*	No function	0		0.3	0	0	0	0
**12*	No function	0.2	0.3	0	0	0	0	0
**13*	No function	0.1		0.3	0.2	0.1	0.1	0
**15*	No function	0.2	0.6	0	0	0	0	0
**20*	No function	0		0	0	0	0	0.1
**21*	No function	0		0	0	0	0	0.5
**31*	No function	0		0	0.2	0.6	0	0
**36*	No function	0.4		0.3	0	0	0	0.2
**36x2*	No function	0		0	0	0	0	0.3
**40*	No function	1.4	1.3	0.3	0	0	0	0
**42*	No function	0.1		0	0	0	0	0
**56*	No function	0.2		0	0	0	0	0
**69*	No function	0.1		0	0.1	0	0.2	0.3
**68*	No function	0		0	0	0.1	0.1	0
**68+*4*	No function	0.1		1	5.7	2.6	2.2	0
**68x2+*4*	No function	0		0	0	0.3	0	0
**99*	No function	0		0	0	0	0.2	0
**1x2*	Increased	0.4	1.1	0.3	0.5	1.2	0.6	0.3
**1x3*	Increased	0.2		0	0	0.1	0	0
**1x2+*83*	Increased	0		0	0	0	0.1	0
**2x2*	Increased	1.9	1.7	1.3	1.5	0.6	0.4	0.5
**2x3*	Increased	0.1		0	0.1	0	0	0.1
**2x4*	Increased	0.2		0	0	0	0	0
**35x2*	Increased	0	0	0	0	0.1	0	0
**45x2*	Increased	0.1		0	0	0	0	0
**45x3*	Increased	0.1		0	0	0	0	0
**1+*90*	Uncertain	0		0	0	0	0	0.1
**22*	Uncertain	0		0	0.3	0	0	0
**28*	Uncertain	0	0	0	0.5	0.1	0	0
**32*	Uncertain	0	0.3	0	0.3	0	0.2	0
**43*	Uncertain	0.8	1.7	1	0.1	0	1	0

Table 5.4 – continued

CYP2D6 allele	CPIC function	Allele frequencies in indicated populations (%)						
		SSA in this study (n=947)	Previous SSA studies [a] (n=2248)	African American/ Afro-Caribbean (n=157)	EUR (n=498)	AMR (n=345)	SAS (n=481)	EAS (n=502)
**43x2*	Uncertain	0.1		0	0	0.1	0	0
**52*	Uncertain	0	0.8	0	0	0	0	0.1
**71*	Uncertain	0		0	0	0	0	0.6
**73*	Unknown	0.2	0.5	0	0	0	0	0
**74*	Unknown	0.1	0.5	0	0	0	0	0
**82*	Unknown	0		0	0	0.4	0	0
**86*	Unknown	0		0	0	0	2.3	0
**106*	Uncertain	1.5		0	0	0.1	0	0
**108*	Unknown	0		0	0.3	0	0	0
**111*	Unknown	0		0	0.4	0	0.8	0
**112*	Unknown	0		0	0	0	0.2	0
**113*	Unknown	0		0	0	0	0.8	0
**117*	Unknown	0		0	0	0	0	0
**119*	Unknown	0		0	0	0	0	0.1
**121*	Unknown	0		0.3	0	0	0	0
**122*	Unknown	0.1		0.3	0	0	0	0
**125*	Unknown	0.3		0.3	0	0	0	0
**139*	Unknown	0.1		0	0	0	0.1	0
**144*	Unknown	0		0	0	0	0	0.1
**149*	Unknown	0.3		0	0	0	0	0.1

n, individuals; SSA; sub-Saharan Africa, African American, African American; EUR, European; SAS, South Asian; EAS, East Asian; CPIC, Clinical Pharmacogenetics Implementation Consortium; PharmGKB, Pharmacogenomics Knowledge Base

[a] See PharmGKB and CPIC *CYP2D6* reference materials (`https://www.pharmgkb.org/page/cyp2d6RefMaterials`)

Table 5.5: *CYP2D6* allele frequencies across all sub-Saharan African populations included in the study.

		Allele frequencies (%)													
***CYP2D6* allele**	**CPIC function**	All (n=947)	Benin (n=49)	BRN (n=49)	Burkina Faso (n=33)	Ghana (n=26)	Cameroon (n=26)	Zambia (n=41)	Botswana (n=47)	South Africa (n=153)	ESN (n=98)	YRI (n=107)	GWD (n=111)	MSL (n=84)	LWK (n=95)
**1*	Normal	24.7	17.3	24.5	30.3	26.9	21.2	32.9	31.9	22.9	29.6	25.7	26.6	18.5	20.5
**2*	Normal	13.4	9.2	15.3	13.6	13.5	21.2	12.2	11.7	13.7	9.2	11.2	14	14.3	18.4
**27*	Normal	0.3	1	0	0	0	0	0	0	0	0	0	0	0	2.1
**34*	Normal	0.1	0	0	0	0	0	0	0	0	0	0	0.5	0	0
**39*	Normal	0.1	1	0	0	0	0	0	0	0	0	0	0	0	0
**45*	Normal	3.7	4.1	3.1	4.5	7.7	3.8	3.7	2.1	4.9	4.6	3.3	1.8	3.6	3.7
**46*	Normal	0.5	0	1	0	0	0	0	1.1	0	1	0	1.4	0	1.1
**17x2*	Normal	0.2	1	0	1.5	0	0	0	0	0	0.5	0	0	0	0
**29x2*	Normal	0.3	2	0	0	0	0	0	0	0	0	0	1.4	0	0.5
**10*	Decreased	3.9	5.1	3.1	1.5	0	1.9	4.9	3.2	4.6	1.5	4.7	5.4	8.9	1.1
**17*	Decreased	19.5	30.6	23.5	15.2	21.2	5.8	14.6	19.1	16.7	22.4	22.9	14.9	23.8	18.4
**29*	Decreased	10.0	4.1	10.2	12.1	5.8	13.5	17.1	6.4	10.1	7.1	9.3	9.9	8.9	16.3
**41*	Decreased	0.8	2	0	0	0	0	0	0	0.7	1.5	0.9	0.5	0.6	2.6
**84*	Decreased	0.1	0	1	0	0	0	0	1.1	0	0	0	0	0	0
**4*	No function	1.7	3.1	0	0	0	3.8	1.2	0	2.3	2.6	0.5	2.7	1.2	2.1
**5*	No function	8.1	7.1	6.1	6.1	7.7	7.7	7.3	13.8	14.7	3.6	6.5	5	8.9	4.7
**12*	No function	0.2	0	0	0	0	1.9	0	2.1	0	0	0	0	0	0
**15*	No function	0.2	0	0	0	0	0	0	0	0	1	0.5	0	0	0
**36*	No function	0.4	0	0	1.5	0	0	0	0	0	0	0	1.8	1.2	0
**40*	No function	1.4	1	1	0	1.9	1.9	0	2.1	2.6	2.6	0.5	0.9	1.8	0
**42*	No function	0.1	0	0	0	0	0	0	0	0	0	0	0.5	0	0.5
**56*	No function	0.2	0	0	0	0	0	2.4	0	0	0	0	0	0.6	0
**69*	No function	0.1	0	0	0	0	0	0	0	0.3	0	0	0	0	0
**68+*4*	No function	0.1	0	0	0	1.9	0	0	0	0	0	0	0	0	0
**13*	No function	0.1	0	0	1.5	0	0	0	0	0	0	0	0	0	0

Table 5.5 – continued

CYP2D6 allele	CPIC function	Allele frequencies (%)													
		All (n=947)	Benin (n=49)	BRN (n=49)	Burkina Faso (n=33)	Ghana (n=26)	Cameroon (n=26)	Zambia (n=41)	Botswana (n=47)	South Africa (n=153)	ESN (n=98)	YRI (n=107)	GWD (n=111)	MSL (n=84)	LWK (n=95)
**4x2*	No function	2.2	2	2	3	3.8	3.8	1.2	0	1.3	5.1	4.7	1.4	1.8	0.5
**4x3*	No function	0.1	0	0	0	0	0	0	0	0	0	0.5	0	0.6	0
**1x2*	Increased	0.4	0	1	0	3.8	0	0	1.1	0	0	0.5	1.4	0	0
**1x3*	Increased	0.2	0	0	0	0	0	0	0	0.7	0	0.5	0	0	0
**2x2*	Increased	1.9	2	3.1	3	0	1.9	0	3.2	0	5.1	2.8	2.3	1.2	1.1
**2x3*	Increased	0.1	0	0	0	0	0	0	0	0	0	0	0	0.6	0
**2x4*	Increased	0.1	2	1	0	0	0	0	0	0	0	0	0	0	0
**45x2*	Increased	0.1	0	0	0	0	1.9	0	0	0	0	0	0	0	0.5
**45x3*	Increased	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0.5
**43*	Uncertain	0.8	0	2	1.5	1.9	0	0	0	0.3	0.5	0	1.8	0	3.2
**43x2*	Uncertain	0.1	0	0	0	0	0	0	0	0	0	0.5	0	0	0
**73*	Unknown	0.2	0	0	0	0	0	0	0	1	0	0	0	0	0
**74*	Unknown	0.1	0	0	0	0	0	0	0	0.3	0	0	0	0	0
**106*	Uncertain	1.5	5.1	0	3	1.9	3.8	1.2	0	0.7	1	3.3	1.8	1.2	0
**122*	Unknown	0.1	0	0	0	0	0	0	0	0	0	0	0.5	0	0
**125*	Unknown	0.3	0	1	0	0	0	0	0	0	0.5	0.5	0.5	0	0.5
**139*	Unknown	0.1	0	0	0	0	0	0	0	0	0.5	0	0	0	0
**149*	Unknown	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0

BRN, Berom in Nigeria; ESN, Esan in Nigeria; GWD, Gambian in Western Division (Mandinka); LWK, Luhya in Webuye (Kenya); MSL, Mende in Sierra Leone; YRI, Yoruba in Nigeria.

5.4.3 CYP2D6 phenotype distribution

The predicted CYP2D6 metaboliser phenotype distributions, translated from the *CYP2D6* diplotype calls using the activity score system, are highlighted in Figure 5.6 and Figure 5.7.

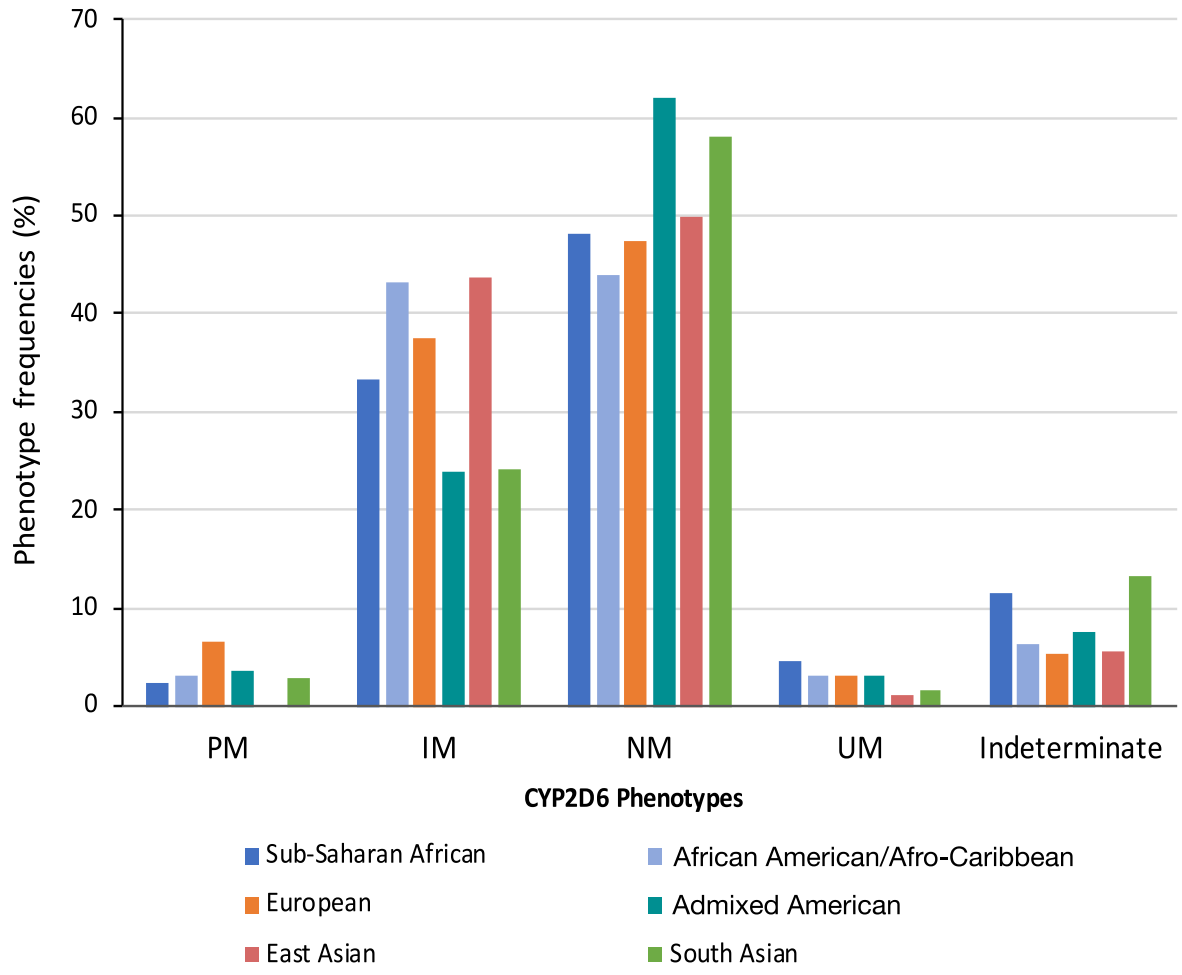


Figure 5.6: Distribution of *CYP2D6* phenotypes predicted from diplotypes called from high-coverage WGS data. PM, poor metaboliser; IM, intermediate metaboliser; NM, normal metaboliser; UM, ultrarapid metaboliser.

Twenty two participants (2.3%) were predicted to be CYP2D6 poor metabolisers (PMs) across the SSA populations in the study (Figure 5.6). There were no significant differences between the PM phenotype distributions across SSA. Each SSA population in the study had at least one CYP2D6 PM except for the BRN, YRI, GHA, and BSZ (Figure 5.7). The highest frequencies of PM diplotypes responsible for the PM phenotype were observed among participants from Botswana (6.4%) and South Africa (3.9%). In comparison, the CYP2D6 PM phenotype was common among the Europeans (6.6%) and no occurrence was observed among the East Asian participants (Figure 5.6).

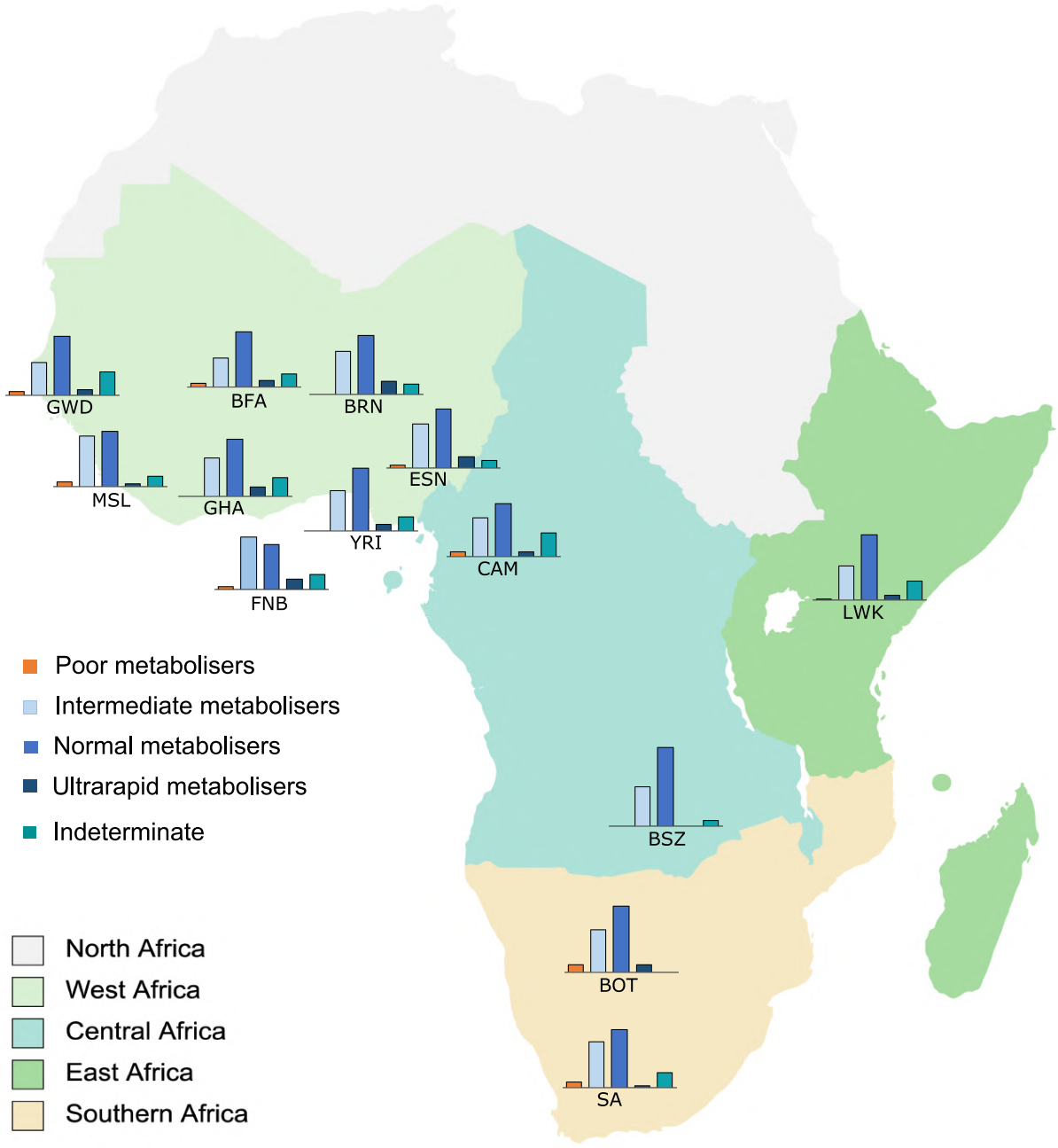


Figure 5.7: CYP2D6 phenotype distribution across the SSA populations in the study. BFA, participants from Burkina Faso; BOT, participants from Botswana; BRN, Berom in Nigeria; BSZ, Bantu-speakers in Zambia; CAM, Cameroonian participants; ESN, Esan in Nigeria; FNB, Fon in Benin; GHA, Ghanaian partcipants; GWD, Gambian in Western Division (Mandinka); LWK, Luhya in Webuye (Kenya); MSL, Mende in Sierra Leone; YRI, Yoruba in Nigeria.

For the intermediate metaboliser (IM) phenotype, considerable differences in proportions were observed across SSA ranging from 26.5% in the GWD to 42% in the FNB (Figure 5.7). This includes a statistically significant difference between the MSL and GWD (p-value = 0.033). The average frequency of diplotypes predicting the IM phenotype across SSA (33.9%) was found to be less than that in African Americans/Afro-Caribbeans (43.3%, p-value = 0.0244), Europeans (37.6%, p-value = 0.1671 i.e. not statistically significant), and East Asians (43.7%, p-value = 0.00027), but significantly higher than the frequency in admixed Americans (23.9%, p-value = 0.0006) and South Asians (24.1%, p-value = 0.000143) (Figure 5.6).

We observed a higher frequency of diplotypes predicting the ultrarapid metaboliser phenotype in SSA (4.7%) compared to that in other biogeographical groups based on the datasets analysed in this study (Figure 5.6). These diplotypes were disproportionately more frequent among participants from western Africa compared to east, central and southern Africa (Figure 5.7), with the highest frequencies being among the Berom (10.2%) and the Esan (9.1%), and the lowest being 0% among the BSZ and 1.3% among the South African participants.

5.4.4 Potential novel *CYP2D6* star alleles inferred from Illumina WGS data

We identified 27 potential novel *CYP2D6* alleles in SSA populations, 19 of which were fully phased (computationally) while 8 were observed in individuals whose diplotypes could not be resolved (Table 5.6). For the phased potential novel alleles, the phasing was made possible either by having the same background allele on both chromosomes or observing multiple participants with the same novel core variant. However, suballeles could not be determined using this approach, so the potential novel alleles are indicated in Table 5.6 using only the core variants. These core variants were all found to be deleterious by at least one of the VEP plugins used in the study, except for rs140900383 (A226V), rs141756339 (R474Q), and rs536109057 (stop gained) (Table 5.7).

Table 5.6: Potential novel African-specific *CYP2D6* alleles inferred from Illumina WGS data

Haplotype	Background allele	Additional core variant(s)	Variant type	Count	Country/dataset[a]
1	*1	rs140900383∼7471C>T	Missense (A226V)	1	Gambia (1000G)
2	*1	rs141756339∼9164G>A	Missense (R474Q)	1	Nigeria
3	*1	rs565013903∼7600G>A	Missense (R269P)	3	Namibia (SGDP)
4	*1	rs567606867∼7004G>A	Missense (E215K)	1	Sierra Leone (1000G)
5	*2	rs368858603∼7609_7610insA	Frameshift	2	Botswana, Kenya (1000G)
6	*2	rs368858603∼7609_7610insA + rs374616348∼6628G>T	Frameshift, Missense (V119L)	2	Cameroon, Congo (SGDP)
7	*2	rs28371704∼6002A>G + rs28371703∼5992C>A	Missense (H94R, L91M)	1	Kenya (1000G)
8	*2	rs375715419∼9115A>G	Missense (T458A)	1	Zambia
9	*2	rs376636053∼6655T>C	Missense (W128R)	1	Sierra Leone (1000G)
10	*2	rs769157652∼8873G>A	Missense (E410K)	1	Sierra Leone (1000G)
11	*17	rs747089665∼8885C>T + rs769157652∼8873G>A	Missense (R414C, E410K)	1	South Africa
12	*17	rs1450231864∼8231T>C	Missense (M347T)	1	South Africa (SGDP)
13	*29	rs76802407∼6012C>G (now *149) [b]	Missense (D97E)	5	Burkina Faso, Ghana Gambia, Sierra Leone (1000G)
14	*29	rs201006451∼6767C>T	Missense (A165V)	1	Nigeria (1000G)
15	*29	rs536109057∼5173C>T	Stop-gained	2	Gambia (1000G)
16	*29	rs760940331∼9096G>A	Missense (M451I)	2	Gambia (1000G), Nigeria (1000G)
17	*41	rs141824015∼8206A>C	Missense (I339L)	4	Kenya (1000G), South Africa
18	*45	rs3915951∼8177G>T	Missense (R329L)	2	Gambia (1000G)
19	*70	rs16947∼7870C>T [c]	Missense (R296C)	5	Cameroon, South Africa

Unresolved sub-Saharan African *CYP2D6* diplotypes with potential novel alleles

#	Background alleles	Additional core variant(s)	Variant type	Number of participants	Country/dataset
1	**1/*43*	rs61731586∼8330G>A	Missense (R380H)	1	Nigeria (1000G)

Table 5.6 – continued

#	Background alleles	Additional core variant(s)	Variant type	Number of participants	Country/dataset[a]
2	**2/*4*	rs61736512∼6679G>A; Possible *CYP2D6* duplication	Missense (V136M)	1	Sierra Leone (1000G)
3	**2/*17*	rs778690377∼6968C>T	Missense (L203F)	1	Kenya (1000G)
4	**2/*17*	rs373243894∼5141C>T	Missense (P41L)	1	Kenya (1000G)
5	**2/*29*	rs556882139∼6879G>A	Missense (R173H)	1	Gambia (1000G)
6	**2/*139*	rs566108360∼8209G>C + rs1135829∼7838A>G; Possible *CYP2D6* duplication	Missense (G340R, N285S)	1	Congo (SGDP)
7	**17/*29*	rs532668079∼9065G>A	Missense (R441H)	1	Nigeria (1000G)
8	**1/*2*	*4, *17, and *29-defining variants in same sample		1	South Africa
9	**1/*2*	Possible *68-like *CYP2D6/2D7* hybrid (lacking 5119C>T)	SV	1	South Africa
10	**1/*2*	Duplicated *CYP2D6* gene copy (*1 + rs565013903∼7600G>A)	SV	1	Namibia
11	**46/*45*	Possible *68-like *CYP2D6/2D7* hybrid (lacking 5119C>T)	SV	1	Gambia (1000G)
12	**29/*45*	Possible *68-like *CYP2D6/2D7* hybrid (lacking 5119C>T)	SV	1	Kenya (1000G)
13	**45/*45*	Possible *68-like *CYP2D6/2D7* hybrid (lacking 5119C>T)	SV	1	Kenya (1000G)

SGDP, Simons Genome Diversity Project; 1000G, 1000 Genomes Project; SV, Structural variant
[a] IDs for Coriell and SGDP samples are provided in Table S5.1 as these samples can potentially be used to reference materials.
[b] This haplotype has recently been validated by (Gaedigk, A. et al., 2021). Subsequent validation is required to determine whether the *CYP2D6*149* suballeles identified in this study are similar to this submission.
[c] Matimba et al. (2009) identified *CYP2D6*70* (moderate evidence on PharmVar) in their study. In our analysis and laboratory validation, we found that individuals with the **70*-defining variants also had 7870C>T (R296C) in phase (subsubsection 5.4.4).

Table 5.7: VEP plugin predictions of the deleteriousness of core variants defining potential novel African-specific *CYP2D6* alleles identified in the study.

Core variants	Consequence	CADD	SIFT	SIFT Indel	Polyphen-2	LRT	PROVEAN	VEST4
rs140900383	missense (A226V)							
rs141756339	missense (R474Q)							
rs565013903	missense (R269P)	X	X		X	X	X	X
rs567606867	missense (E215K)							X
rs368858603	frameshift			X				
rs374616348	missense (V119L)							
rs28371704	missense (H94R)						X	
rs28371703	missense (L91M)	X	X		X	X		
rs375715419	missense (T458A)	X	X			X	X	
rs376636053	missense (W128R)	X	X		X		X	X
rs769157652	missense (E410K)	X						
rs747089665	missense (R414C)	X					X	
rs1450231864	missense (M347T)	X			X	X	X	X
rs201006451	missense (A165V)							X
rs536109057	stop-gained							
rs760940331	missense (M451I)	X	X		X	X	X	X
rs141824015	missense (I339L)				X	X		X
rs3915951	missense (R329L)	X	X					
rs61731586	missense (R380H)	X				X	X	X
rs61736512	missense (V136I)							
rs778690377	missense (L203F)				X			
rs373243894	missense (P41L)	X	X		X	X	X	X
rs556882139	missense (R173H)	X	X					
rs532668079	missense (R441H)	X	X		X	X	X	X
rs566108360	missense (G340R)	X	X		X	X	X	X
rs1135829	missense (N285S)						X	

X, indicates that the variant was predicted to be deleterious by the corresponding plugin; CADD, Combined Annotation Dependent Depletion; SIFT, Sorting Intolerant from Tolerant; Polyphen-2, Polymorphism Phenotyping (version 2); LRT, Likelihood Ratio Test; PROVEAN, Protein Variation Effect Analyzer; VEST4, Variant Effect Scoring Tool (version 4). For each plugin a variant was considered deleterious if the scores met the following ADME-optimised thresholds suggested by Zhou et al. (2017): CADD PHRED (> 19.19), SIFT (< 0.0376), Polyphen-2 (> 0.3841), LRT (< 0.0025), PROVEAN (< -3.28), VEST4 (> 0.4534).

5.4.5 *CYP2D6* novel allele characterisation using targeted SMRT sequencing

CYP2D6 XL-PCR was used to generate fragments for long-read sequencing involving 190 participants, 32 from Burkina Faso, 26 from Ghana, and 132 from South African cohorts (AWI-Gen and CBRL). For one participant who was predicted to have *CYP2D6*13/*45*, the **13* CYP2D7-2D6 hybrid fragment did not amplify — we could not resolve whether the issue related to sample quality or absence of *CYP2D6*13*. For all the other participants, the corresponding XL-PCR *CYP2D6* fragments were successfully amplified in preparation for subsequent long-read sequencing.

From the first two batches of samples (n=127) sent for *CYP2D6* HiFi sequencing (along with *CYP2A6* and *CYP2B6* XL-PCR fragments), *CYP2D6* XL-PCR fragments were successfully barcoded in only 82 of these participants. Generally, the *CYP2D6* diplotypes derived from targeted PacBio HiFi resequencing were concordant to those inferred from short-read-based calls across all the samples included in the laboratory validation study.

HiFi sequencing enabled the characterisation of 2 predicted novel major alleles (Figure 5.8) in this batch of samples and revealed novel suballeles in multiple participants.

- **Case 1**: Referring back to the short-read-based allele calling, StellarPGx and the other algorithms used could not assign a diplotype to 3 of the individuals carrying *CYP2D6*70* core SNVs, while 2 other participants had been genotyped as *CYP2D6*34/*70*. However, our HiFi data analysis for one of the South African participants revealed that the *CYP2D6*2* SNP (rs16947, NG_008376.4:g.7870C>T, R296C) was in phase with the current *CYP2D6*70* allele-defining variants (Figure 5.8). This participant had *CYP2D6*5* (gene deletion) as the second allele in the diplotype.
- **Case 2**: The second validated novel major allele (Haplotype 17, Table 5.6) was also found in a South African participant with *CYP2D6*5* as the second allele in the diplotype. HiFi sequencing showed that the rs141824015 missense variant (NG_008376.4:g.8206A>C, I339L) was in phase with the *CYP2D6*41* defining variant rs28371725 (NG_008376.4: g.8008G>A, splice defect) on a **2* background (Figure 5.8).
- **Suballeles**: Among the novel suballeles, we ascertained the presence of subvariants not yet catalogued by PharmVar for *CYP2D6*149* and *CYP2D6*36* (Figure 5.9).

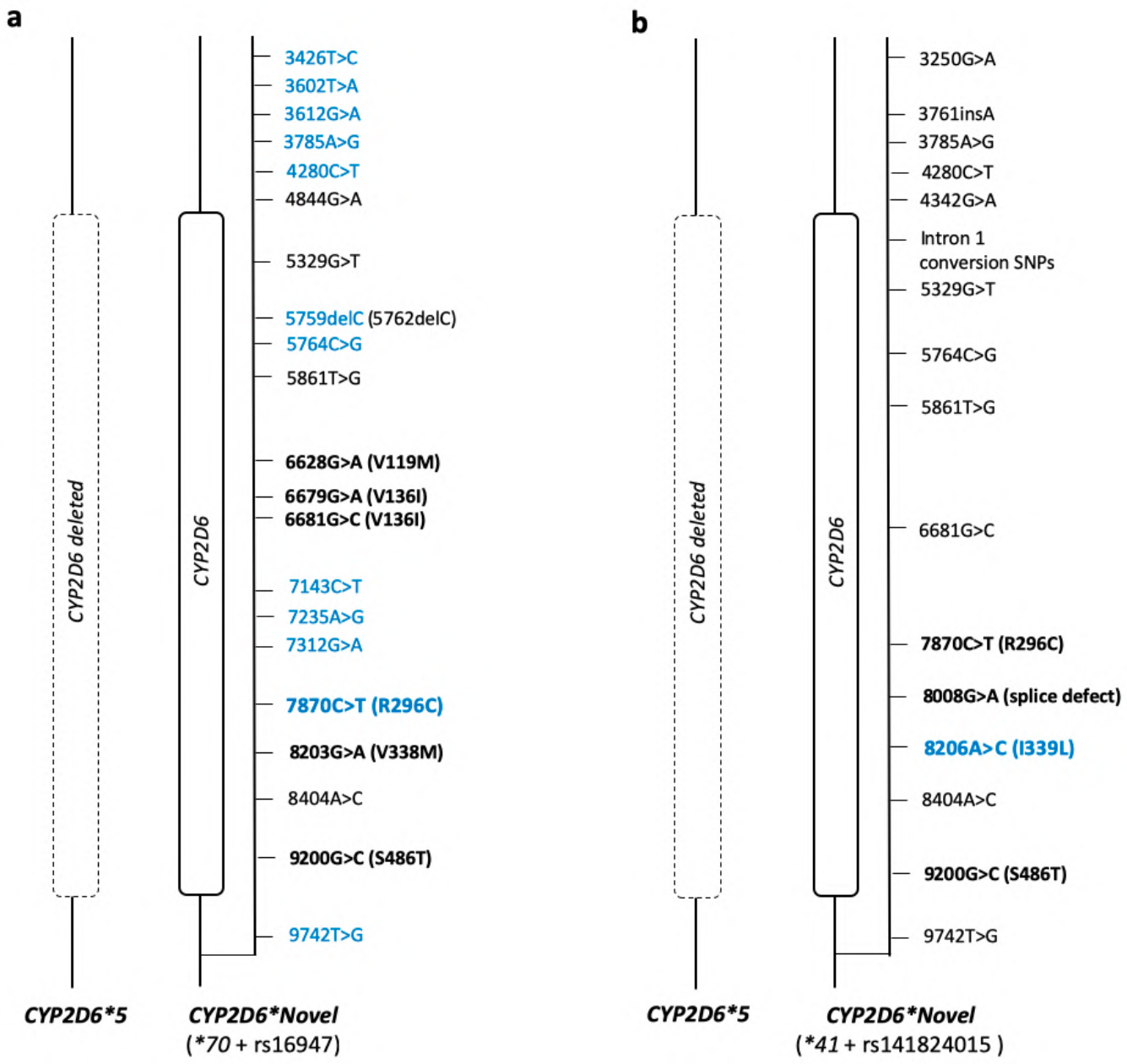


Figure 5.8: Novel *CYP2D6* star alleles in SSA characterised via XL-PCR and HiFi sequencing. Panel (**a**) shows the two haplotypes in a South African participant with the novel *CYP2D6*70* + rs16947 (R296C) allele. The submission of this allele for consideration by PharmVar is ongoing. The variant phasing and frequency information from our analysis suggests that *CYP2D6*70*, which has a 'moderate' PharmVar level of evidence, should be re-defined to include rs16947 (NG_008376.4:g.7870C>T, R296C) but the final decision on this rests with the PharmVar *CYP2D6* expert panel. Panel (**b**) shows the two alleles identified in a South African participant with the novel *CYP2D6*41* + rs141824015 (I339L) allele. Two suballeles of this haplotype have been predicted in the Luhya in Webuye, Kenya (LWK) from the 1000 Genomes Project dataset and they are being charaterised in a collaborative project with the PharmVar team.

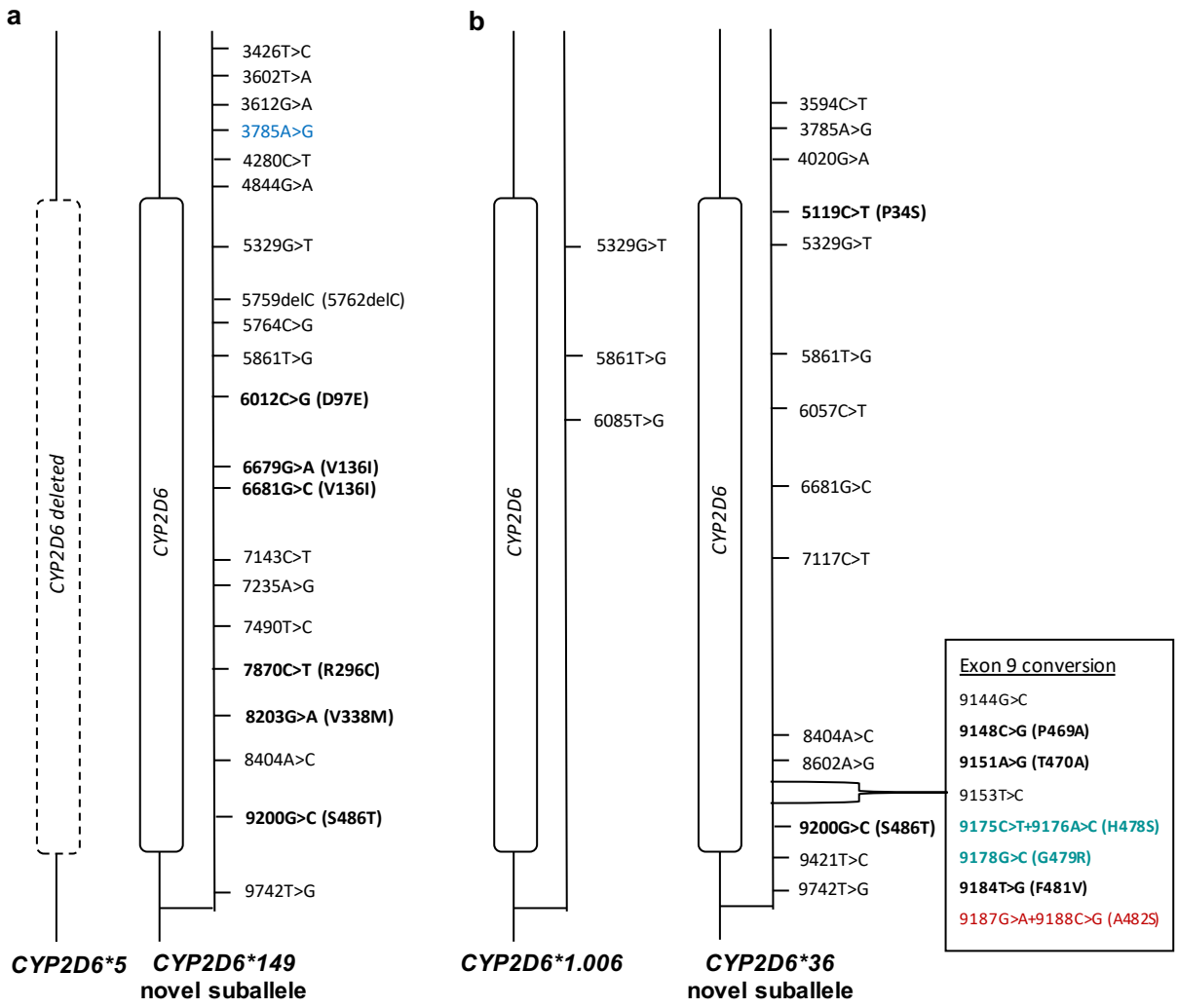


Figure 5.9: Novel *CYP2D6* suballeles in SSA characterised via XL-PCR and HiFi sequencing. Panel (**a**) shows a novel *CYP2D6*149* suballele from this study. *CYP2D6*149* is defined by rs76802407 (D97E) in phase with the **29*-defining variants, and has recently been reviewed and assigned a star number by PharmVar (`https://www.pharmvar.org/gene/CYP2D6`). The allele depicted here differs from *CYP2D6*149.001* as it has the NG_008376.4:g.3785A>G variant. Panel (**b**) shows a novel *CYP2B6*36* suballele called in a participant from Burkina Faso. HiFi data facilitated the detection of the exon 9 conversion SNVs. The SNVs indicated in the red text are known to be part of this gene conversion (see `https://www.pharmvar.org`) but we did not detect them in the HiFi data for this participant, while SNVs indicated in teal had very low read support.

5.5 Discussion

African populations have so far been understudied for *CYP2D6* pharmacogenomic variation in comparison to European and Asian populations (Gaedigk et al., 2017). This study represents one of the most comprehensive investigations of the *CYP2D6* allelic variation across SSA to date. We analysed 962 high-depth African genomes generated mainly by H3Africa projects other collaborative projects within Africa, and the 1000 Genomes Project Consortium for common, rare and novel *CYP2D6* star alleles, including structural variants. In addition, we assessed the distribution of CYP2D6 metaboliser phenotypes predicted from participant diplotypes across SSA, and also performed laboratory validation for some of the novel African-specific alleles via XL-PCR and SMRT sequencing (Pacific Biosciences). Furthermore, we analysed 1000 Genomes Project-generated high-depth genomes from other biogeographical groups for comparison of the *CYP2D6* allele and phenotype distributions in these populations with the allelic landscape observed in SSA.

The known *CYP2D6* allele distributions in SSA populations are from studies among participants from Ethiopia (Ahmed et al., 2019; Aklillu et al., 1996),Kenya (Gutiérrez Rico et al., 2020), Tanzania (Bathum et al., 1999; Dandara et al., 2001; Wennerholm et al., 2001), Zimbabwe (Dandara et al., 2001; Matimba et al., 2009; Masimirembwa et al., 1996), South Africa (Awandu et al., 2018; Wright et al., 2010), and Ghana (Griese et al., 1999). Most of these studies were relatively small-scaled by design thus limiting the capacity for making allele distribution comparisons across diverse SSA populations and also for detecting novel alleles. In this study, we present the *CYP2D6* allele distributions and predicted novel alleles from understudied populations for whom high coverage WGS data has recently been generated e.g. the Berom in Nigeria, Fon in Benin, and participants from Burkina Faso, Botswana, Zambia, Cameroon, and South Africa (Choudhury et al., 2020; da Rocha, Othman, Botha, et al., 2021). In addition, our study included well-studied populations (GWD, MSL, ESN, LWK, YRI) where previous lack of high coverage WGS data had made it virtually impossible to characterised the allelic variation in the hypervariable *CYP2D* region (Drögemöller, Wright, et al., 2013).

Our *CYP2D6* star allele analysis reveals largely distinct allele distributions across the different SSA populations in this study (Table 5.5), including differences between populations in neighbouring countries and/or ethnolinguistic groups within the same country (e.g. the Berom, Esan and Yoruba in Nigeria). However, the estimated frequencies of the rarer star alleles should be interpreted with caution given the relatively small sample sizes in the SSA populations included in this study. The frequencies of the key African-specific decreased function *CYP2D6* alleles (i.e. **17* and **29*) are consistent with previously reported frequencies from studies within SSA (Dodgen et al., 2013; Matimba et al., 2009; Wright et al., 2010). However, they are by no means uniform across the whole of SSA (Table 5.4). Conversely, the frequencies of some other

key alleles such as *CYP2D6*5* and *CYP2D6*41* in SSA were significantly different from the average frequency estimated by the PharmGKB from previous studies in the region (Table 5.4). Furthermore, some of the *CYP2D6* alleles e.g. **4xN* (no function), **36* (no function) **106* (uncertain function), **125* (unknown function), **1xN* (increased function), **2xN* (increased function), not included in the SSA catalogue in the current PharmGKB and CPIC *CYP2D6* reference materials were identified in SSA in this study. There were some allele frequency differences observed between the SSA populations and the African Americans/Afro-Caribbeans included in the study. For example, *CYP2D6*29* was found to be less frequent in the African Americans/Afro-Caribbeans compared to the SSA populations in the study, except for the FNB, GHA and BOT participants where it was comparable (Table 5.4 and Table 5.5), while *CYP2D6*4* (no function) was more frequent in the African Americans/Afro-Caribbeans as compared to SSA. Moreover, the differences in allele frequencies observed between SSA and European, East Asian and South Asian populations (Table 5.4) underline the need for more comprehensive genotyping panels capable of assaying the extensive *CYP2D6* allelic diversity in the different biogeographical groups.

We observed unique distributions of the CYP2D6 metaboliser phenotypes across different SSA populations (Figure 5.7) and also in comparison to African American/Afro-Caribbean, European, East Asian and South Asian populations (Figure 5.6) as a result of differences in the diplotype frequencies and in the distribution of key alleles (Figure 5.5). For example, the MSL had a statistically significant higher proportion of CYP2D6 intermediate metabolisers compared to the GWD. In addition, some CYP2D6 metaboliser phenotypes were not observed in some of the populations e.g. the YRI (n=108), BRN (n=49) and BSZ (n=41) had no predicted CYP2D6 poor metabolisers. However, the relatively low sample sizes among the SSA populations our study made it difficult to ascertain whether some of the observed differences in CYP2D6 phenotype distribution were statistically significant. This emphasises the need for extensive allele characterisation and phenotypic studies across Africa in order to develop effective precision medicine strategies on the continent, in particular for medications metabolised by CYP2D6.

Twenty seven potential novel *CYP2D6* star alleles were observed among the SSA populations included in this study given the high depth of the WGS data analysed. The core variants defining these alleles were all rare and are not novel per se but are rather nonsynonymous variants that are either not currently catalogued as allele-defining by PharmVar or have been found in different combinations in our study and evaluated based on the PharmVar criteria. The 'novel' combinations of known core variants identified in our study represent allelic information that could be missed if comprehensive haplotyping is not done e.g. when relying on genotyping arrays and imputation, and they could have implications for the CPIC function of the relevant *CYP2D6* haplotypes. Given the genotyping challenges associated with *CYP2D6*, computationally-inferred novel alleles should be interpreted with caution. Through high fidelity (HiFi) long-read

sequencing (Pacific Biosciences), we characterised two of the novel *CYP2D6* star alleles and also found multiple novel suballeles (Figure 5.8 and Figure 5.9). The process of submitting these novel alleles to PharmVar is ongoing. The presence of these rare previously uncharacterised alleles could be responsible for some of the observed differences in the response among Africans to medications metabolised by *CYP2D6*.

Our study had some limitations. Firstly, we focused our analysis on SSA and did not analyse any northern African genomes except for one Algerian participant (LP6005441-DNA_G08, SGDP dataset). Furthermore, this study included very few samples from Nilo-Saharan and Afroasiatic language speakers. We also had a limited number of samples from speakers of non-Bantu languages SSA such as the Khoe and San speakers. As demonstrated in this study, more diverse sampling is needed to obtain a more comprehensive landscape of variation in *CYP2D6* and other key pharmacogenes. Nonetheless, this study has revealed key insights regarding the *CYP2D6* pharmaco-genomic variation in the African populations included in the analysis and, importantly, established analysis pipelines that can be easily implemented and adapted for future work involving unsampled populations across Africa.

Secondly, we were unable to computationally resolve 13 *CYP2D6* diplotypes in SSA either due to presence of novel core variants that could not be phased or due to potential uncharacterised structural variations (Table 5.6) that were novel to all the algorithms used in the star allele calling. Nonetheless, this study has summarised Coriell sample IDs of participants with these diplotypes so that they be included in follow-up studies to this work.

Thirdly, due to the occurrence of existing alleles with unknown/uncertain function and multiple novel alleles, the *CYP2D6* phenotype for 10% of the study participants in SSA was indeterminate as *in vitro* and *in vivo* studies were out of the scope of this work. This remains one of the main challenges of using the activity score system for translating *CYP2D6* genotype to phenotype. In addition, there are a number of factors that could influence the CYP2D6 phenotype e.g. phenoconversion which we did not investigate due to insufficient metadata. Furthermore, the deleterious predictions by the VEP plugins used in the study for core variants defining potential novel alleles may not necessarily correlate with the actual CPIC function.

Fourthly, in our laboratory validation, *CYP2D6* XL-PCR products for 30 participants failed to barcode during the HiFi sequencing library preparation due to technical challenges that arose from pooling the *CYP2D6* amplicons with those from *CYP2A6* and *CYP2B6*. We mitigated this by doing barcoding for select amplicons from samples that had predicted novel alleles or suballeles.

In conclusion, the differences in *CYP2D6* allele frequencies observed across SSA cohorts and in comparison to other global populations underline the need for precision medicine across Africa. In particular, the high number of potential novel *CYP2D6* alleles observed in SSA emphasise the

need for in-depth pharmacogene allele characterisation across Africa, especially in understudied ethnolinguistic groups. We recommend the consensus *CYP2D6* star allele calling approach used in this study for similar analyses focusing on African and other global populations especially when researchers have high-coverage WGS data available to them. Future work entailing the unequivocal characterisation of the novel alleles not validated in this study as well as functional studies to determine their clinical function will be critical in supporting clinical pharmacogenomics implementation strategies across Africa.

5.6 Data Availability Statement

The 1000 Genomes data and the SGDP data are publicly available. The HAAD dataset is available from European Genome-Phenome Archive (`https://ega-archive.org/`) on application to the relevant Data Access Committees (EGADs: EGAD00001003791, EGAD00001006418, EGAD00001004220, EGAD00001004448, EGAD00001004505, EGAD00001004533, EGAD00001004557, EGAD00001004393, EGAD00001007589).

5.7 Author contributions

D.T., Z.L., and S.H. designed the research; S.H. obtained funding, administrative, technical or material support; D.T. analysed the data; D.T. drafted the manuscript; Z.L. and S.H. critically revised the manuscript for important intellectual content; Z.L. and S.H. supervised the research.

Supplementary note

The H3Africa AWI-Gen study (provided data from Ghana, Burkina Faso and South Africa) was approved by the Wits HREC-Medical (protocol numbers M121029 and M170880). The (South Africa) data from the Cell Biology Research Laboratory (CBRL), National Institute for Communicable Diseases (NICD, Wits) was generated by a study approved by the Wits HREC-Medical (protocol number M140926). The SAHGP study was approved by the Wits HREC-Medical (protocol number M120223). The H3Africa Benin study was approved by the Comité d'éthique de la recherche, Université de Montréal. The H3Africa CAfGEN study (provided data from Botswana) was approved by the Institutional Review Board (IRB) of the Ministry of Health of the Republic of Botswana (PPPME-13/18/1). The H3Africa TrypanoGEN Study (Cameroon component) was approved by the Comité National D'Ethique de la Recherche pour la Santé Humaine of the Republic of Cameroon (2013/11/364/L/CNERSH/SP). The H3Africa TrypanoGEN Study (Zambian component) was approved by the Biomedical Research Ethics Committee of the University of Zambia (FWA00000338). The H3Africa ACCME Study (Nigeria) was approved by the National Health Research Ethics Committee of Nigeria (NHREC/01/01/2007-29/11/2016).

Table S5.1: Sample information for participants with publicly available high coverage WGS data containing potential novel *CYP2D6* star alleles. The Coriell DNA samples from these participants could potentially be added to GeT-RM reference materials for *CYP2D6* genotyping in pharmacogenomics testing laboratories.

Coriell / SGDP IDs	Predicted consensus diplotype	Country/ population	Primary project
HG02570	*1/[*1+rs140900383∼7471C>T(A226V)]	GWD	1000G
LP6005441-DNA_A11	[*2/*1 + rs565013903∼7600G>A(R269P)]; (cn=3)	Namibia	SGDP
LP6005443-DNA_G08	[*1+rs565013903∼7600G>A]/[*1+rs565013903∼7600G>A]	Namibia	SGDP
HG03428	*5/[*1+rs567606867∼7004G>A(E215K)]	MSL	1000G
NA19468	*1/[*2+rs368858603∼7609_7610insA (fs)]	LWK	1000G
SS6004471	*2/[*2 + rs368858603∼7609_7610insA (fs) + rs374616348∼6628G>T(V119L)]	Congo	SGDP
NA19314	*5/[*2 + rs28371704∼6002A>G(H94R) + rs28371703∼5992C>A(L91M)]	LWK	1000G
HG03559	*2/[*2+rs376636053∼6655T>C(W128R)]	MSL	1000G
HG03469	*2/[*2+rs769157652∼8873G>A(E410K)]	MSL	1000G
LP6005592-DNA_C05	*5/[*17+rs1450231864∼8231T>C(M347T)]	South Africa	SGDP
HG02666	*1/[*29 + rs76802407∼6012C>G(D97E)]	GWD	1000G
HG02870	*1/[*29+rs76802407∼6012C>G(D97E)]	GWD	1000G
HG03442	*2x2/[*29 + rs76802407∼6012C>G(D97E)]	MSL	1000G
NA18933	*5/[*29+rs201006451∼6767C>T(A165V)]	YRI	1000G
HG02840	*17/[*29+rs536109057∼5173C>T(stop-gained)]	GWD	1000G
HG02860	*2/[*29+rs536109057∼5173C>T(stop-gained)]	GWD	1000G
NA19130	*106/[*29+rs760940331∼9096G>A(M451I)]	YRI	1000G
HG02807	*17/[*29+rs760940331∼9096G>A(M451I)]	GWD	1000G
NA19316	[*2/[*41 + rs141824015∼8206A>C(I339L)]	LWK	1000G
HG02623	*2/[*45 + rs3915951∼8177G>T(R329L)]	GWD	1000G
HG02642	*17/[*45 + rs3915951∼8177G>T]	GWD	1000G
NA19222	[*1/*43] + rs61731586∼8330G>A (R380H)	YRI	1000G
HG03470	[*2/*4] + rs61736512∼6679G>A(V136M); cn=3	MSL	1000G
NA19472	[*2/*17] + rs778690377∼6968C>T (L203F)	LWK	1000G
NA19026	[*2/*17] + rs373243894∼5141C>T(P41L)	LWK	1000G
HG02852	[*2/*29] + rs556882139∼6879G>A(R173H)	GWD	1000G
LP6005441-DNA_A08	[*2/*139] + rs566108360∼8209G>C(G340R) + rs1135829∼7838A>G(N285S)	Congo	SGDP
HG03313	[*17/*29] + rs532668079∼9065G>A(R441H)	ESN	1000G
HG02621	[*45/*46] + Possible *68-like *CYP2D6/2D7* hybrid (lacking 5119C>T)	GWD	1000G
NA19017	[*29/*45] + Possible *68-like *CYP2D6/2D7* hybrid (lacking 5119C>T)	LWK	1000G
NA19456	[*45/*45] + Possible *68-like *CYP2D6/2D7* hybrid (lacking 5119C>T)	LWK	1000G

cn, copy number; fs, frameshift; 1000G, 1000 Human Genomes Project; SGDP, Simons Genome Diversity Project; ESN, Esan in Nigeria; LWK, Luhya in Webuye Kenya; GWD, Gambian from Western Divisions (Mandinka); MSL, Mende in Sierra Leone; YRI, Yoruba in Ibadan, Nigeria; GeT-RM, Genetic Testing Reference Materials Coordination Program.

Chapter 6

Characterisation of the *CYP2B6* and *CYP2A6* allelic variation in African populations

Manuscript *in preparation* by David Twesigomwe, Zané Lombard and Scott Hazelhurst. See subsection 6.6 for author contributions.

6.1 Summary

Background: Genetic variations in *CYP2B6* and *CYP2A6* are known to have an impact on the interindividual response to efavirenz and nicotine. However, the full catalogue of pharmacogenetic variants in these genes is yet to be established, especially for African populations. **Methods**: This study analysed 961 high-depth African genomes to examine the distribution of *CYP2B6* and *CYP2A6* star alleles across sub-Saharan Africa. Novel African-specific *CYP2B6* and *CYP2A6* alleles in a subset of the study population were subsequently characterised via targeted long-read resequencing. **Results**: Non-uniform frequency distributions of *CYP2B6* and *CYP2A6* alleles and predicted phenotypes (for CYP2B6) were observed across different African ethnolinguistic groups. In addition, 50 novel African-specific alleles were computationally predicted in *CYP2B6* and *CYP2A6* combined. Two novel alleles in each gene and several suballeles were further characterised using targeted high fidelity SMRT sequencing. **Conclusion**: The *CYP2B6* and *CYP2A6* allele distributions, and the novel alleles, observed across the various African populations are important for the design of comprehensive genotyping panels, and they could inform personalised medicine strategies relating to safe and effective clinical use of efavirenz and smoking cessation treatment. More broadly, the results of this study highlight the importance of sampling diverse African ethnolinguistic groups for accurate characterisation of the pharmacogene variation landscape across the continent.

6.2 Introduction

Genetic variation in the cytochrome P450 (CYP) supergene family is a major contributor to the drug response variability within and between populations. Of the 57 known functional CYP genes, 12 encode enzymes responsible for the metabolism and bioactivation of 70-80% of all clinically prescribed medications. CYP genetic variants that could impact drug response include various combinations of small nucleotide variations (SNVs) and structural variations including copy number variations (CNVs) and other more complex rearrangements. However, the full catalogue of pharmacogenomically relevant haplotypes ("star alleles") is yet to be determined in African populations and other understudied biogeographical groups (`https://www.pharmgkb.org/page/cyp2b6RefMaterials`). The paucity of information on *CYP* allele frequency distributions in these populations poses challenges for effective application of genotyping arrays and gene panels, understanding the variability observed within activity score categories, and effectively hampers efforts aimed at bringing personalised medicine to reality.

The CYP2B6 enzyme is involved in the bioactivation of 8-13% of drugs in clinical use and is especially central to the metabolism of efavirenz, one of the key drugs for the treatment of HIV type-1 infection. *CYP2B6* variants such as *CYP2B6*6* and *CYP2B6*22* have been shown to account for part of the interindividual variability in efavirenz response (Maimbo et al., 2012). Therefore, considering the *CYP2B6* genotype and the predicted metaboliser status before administering efavirenz to patients is essential in order to minimise the risk of efavirenz toxicity (e.g. neurological toxicity and hepatic injury). CYP2B6 is also involved in the metabolism of nicotine and bupropion. Therefore, *CYP2B6* genetic variations can also have implications for smoking cessation drug therapy and treatment of major depressive disorders (Zhu et al., 2012).

CYP2A6 is the predominant enzyme involved in the metabolism of nicotine, the main psychoactive substance in cigarette smoke (Messina and Tyndale, 1997). *CYP2A6* genetic variation contributes to interindividual differences in nicotine metabolism and it also has an impact on smoking cessation therapy (Tanner and Tyndale, 2017). To a lesser extent, *CYP2A6* variation also contributes to the variability in antiretroviral drug metabolism and response (Ogburn et al., 2010; Soeria-Atmadja et al., 2012).

The human *CYP2B6* and *CYP2A6* genes are located on chromosome 19q13 within the large *CYP2ABFGST* gene cluster. Both genes are in close proximity with homologous pseudogenes, *CYP2B7* and *CYP2A7* respectively. *CYP2B6* and *CYP2A6* are highly polymorphic, with over 37 and 45 star alleles catalogued for these genes, respectively, by the Pharmacogene Variation Consortium (`https://www.pharmvar.org`). The presence of the aforementioned pseudogenes and complex structural variant alleles make genotyping *CYP2B6* (Desta et al., 2021) and *CYP2A6* (Wassenaar et al., 2016) challenging and oftentimes labor-intensive. However,

the decreasing cost of HTS data generation and availability of recently developed star allele calling bioinformatics tools (S.-B. Lee, Wheeler, Thummel, et al., 2019; Numanagić et al., 2018) provide the opportunity to investigate known alleles and to identify potential novel allelic arrangements that can be prioritised for experimental validation.

As shown in previous pharmacogenetics studies, individuals of African ancestry are more likely to have *CYP2B6* decreased- or no-function alleles and poor metaboliser phenotype as regards to efavirenz therapy (`https://www.pharmgkb.org/page/cyp2b6RefMaterials`). For example, the *CYP2B6*6* allele (decreased function) occurs at a frequency of 37% in sub-Saharan African populations, which is slightly higher than that reported for African Americans/ Afro-Caribbeans (32%) but significantly higher than the frequency in Europeans (23%). In the same vein, some key *CYP2A6* star alleles known to cause decreased CYP2A6 enzyme activity are more prevalent in African populations. These include *CYP2A6*17*, **28* and **35* (Tanner and Tyndale, 2017). However, the distribution of known actionable *CYP2B6* and *CYP2A6* star alleles has been investigated in only a few African ethnolinguistic groups. Moreover, previously unsampled African ethnolinguistic groups might habour novel African-specific *CYP2B6* and *CYP2A6* alleles (including structural variants) of functional impact.

This study therefore aims to investigate the distribution of known and potential novel *CYP2B6* and *CYP2A6* star alleles across diverse African populations, and to predict the phenotype distribution where possible. High coverage genomes from other biogeographical groups were included for comparative analyses. Furthermore, we characterise some of the novel *CYP2B6* and *CYP2A6* alleles and suballeles using targeted single molecule real-time (SMRT) sequencing. Although SMRT sequencing-based allele characterisation has been validated for *CYP2D6* (Buermans et al., 2017; Liau et al., 2019; Qiao et al., 2016), similar efforts are lacking for *CYP2B6* and *CYP2A6*. Our targeted HiFi resequencing therefore contributes to addressing this gap in *CYP2B6* and *CYP2A6* allele validation methodology.

Our assessment of the distribution of *CYP2B6* and *CYP2A6* star alleles, and novel allele characterisation provides information that could guide future genotyping strategies across Africa. In particular, this study emphasises the need to study more unsampled populations to obtain a more accurate landscape of *CYP2B6* and *CYP2A6* allele and phenotype diversity, which could subsequently enhance the effectiveness of clinical pharmacogenetics implementation strategies across Africa.

6.3 Materials and Methods

6.3.1 Ethics statement and study datasets

This study was performed under the same ethical clearance (Appendix B on page 215) obtained for the research described in chapter 5. The WGS datasets used in this study representing SSA (n=961) and global (for comparison) populations are the same as those used in Chapter 5 (see Figure 5.1 on page 115 and Table 5.1 on page 116).

6.3.2 Star allele analysis

CYP2B6 and *CYP2A6* star alleles were called from the WGS samples using three separate tools including StellarPGx (v1.2.6) (see Chapter 4), Aldy (v3.1) (Numanagić et al., 2018), and Stargazer (v1.0.8) (S.-B. Lee, Wheeler, Thummel, et al., 2019). The run details and computational resources are similar to those used for the *CYP2D6* allele analysis in subsubsection 5.3.3.

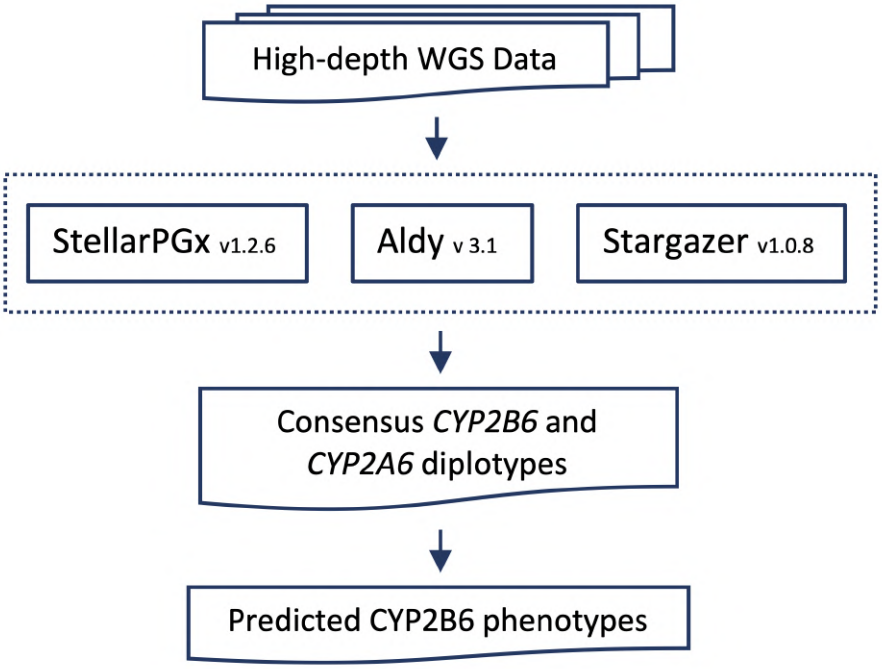


Figure 6.1: Overview of the *CYP2B6* and *CYP2A6* star allele calling workflow used in the study.

6.3.3 Metaboliser phenotype prediction

The *CYP2B6* consensus diplotype calls were translated to CYP2B6 metaboliser phenotypes based on the CPIC guidelines for efavirenz (Desta et al., 2019). However, CYP2A6 phenotypes were not predicted in this study (see subsection 6.5).

6.3.4 Targeted long-read resequencing

6.3.4.1 *CYP2B6* and *CYP2A6* long-range PCR

As *CYP2B6* is a large gene (∼27 kb), multiple XL-PCR fragments fragments were generated to cover various regions (see Figure S6.1 on page 178). *CYP2B6* Frag1 (∼ 9 kb) stretched from the *CYP2B6* upstream region to intron 1; Frag3 (∼ 8.5 kb) covered exon 2 to exon 6; Frag4 (∼ 10.3 kb) covered exon 4 to exon 9; and Frag5 (∼7.1 kb) covered exon 8 to the *CYP2B6* downstream region. We ran multiple PCR optimisations to amplify Frag2 (∼13 kb, exon 1 to exon 3) but none were successful. The XL-PCR primers and cycling conditions for each fragment are detailed in Table 6.1 and Table 6.2.

Table 6.1: Primers used for generating *CYP2B6* and *CYP2A6* long-range PCR fragments in this study

Assay	Primer name	Direction	5'–sequence	References
CYP2B6 Frag1 (∼9.1 kb)	2B6 5Pr F	F	CCTACAAGAAACACATCTCACCAGGA	(Hesse et al., 2004)
	2B6 int1 R	R	AGGAGCTGGAAGGTATTGTCAGTG	
CYP2B6 Frag3 (∼8.5 kb)	2B6 ex2 F/2	F	GCTTTACTGGCCCAACCAGA	(Lang et al., 2004)
	2B6 int6 R	R	TGGTTGGAATCTAGCCCAGAGC	(Lang et al., 2004)
CYP2B6 Frag4 (∼10.3 kb)	2B6 ex4 F	F	CTCGGTCTGCCCATCTATAAAC	(Lang et al., 2004)
	2B6 ex9 R	R	CCTGCACTCACTTGCAATGT	(Twist et al., 2016)
CYP2B6 Frag5 (7-8 kb)	2B6 int7 F	F	TGGAGTGTGTGGAGGGTTGG	(Lang et al., 2004)
	2B6 3Pr R	R	CCTGTGAGCTGTTTCCCGTCT	
CYP2B6 FragH1 (10.3 kb)	2B6 ex4 F	F	CTCGGTCTGCCCATCTATAAAC	(Martis et al., 2013)
	2B7 ex9 R	R	TCTGGTTTCAGAGCAGCTTCC	(Martis et al., 2013)
CYP2B6 FragH2 (8-9 kb)	2B6 ex2 F/2	F	GCTTTACTGGCCCAACCAGA	(Martis et al., 2013)
	2B7 int5 R	R	CTCCCTCTGTCTTTCATTCTGTC	(Martis et al., 2013)
CYP2A6 FragA (9.2 kb)	2A65Pr1F	F	ACCTAGACTTAATCTTCCCGTATAC	(Wassenaar et al., 2016)
	2A6R13	R	GCCTCCCATAGTGCTATAATTAACA	(Wassenaar et al., 2016)
CYP2A6 FragD (6-8 kb)	2A6F3	F	TAGACAGATTCTTAAAAAGCACCT	(Wassenaar et al., 2016)
	2A6Rdup	R	AATTCCTGGATTGACAAGAG	(Oscarson et al., 2002)
CYP2A7-2A6 hybrids (6-9 kb)	2A75Pr1F	F	ACCTAGACTTAATCTTCCCATATAT	(Wassenaar et al., 2016)
	2A6R13	R	GCCTCCCATAGTGCTATAATTAACA	(Wassenaar et al., 2016)

F, Forward; R, Reverse; ex, exon; int, intron; Frag, Fragment; H, Hybrid fragment

Table 6.2: PCR cycling conditions for generating *CYP2B6*, and *CYP2A6* fragments

	CYP2B6 and *CYP2A6* assays									
	1	2	3	4	5	6	7	8	9	10
First denaturation (ºC:s)	94:180									
Denaturation (ºC:s)	94:20	94:20	94:20	94:20	94:20	94:20	94:20	94:20	94:20	94:20
Annealing (ºC:s)	61:30	61:30	61:30	61:30	60:30	61:30	58:30	53:45	51:45	51:45
Extension (ºC:s)	65:660	65:660	65:780	65:660	65:780	65:660	65:660	65:600	65:660	65:660
Extension (ºC:s)	65:660	65:660	65:780	65:660	65:780	65:660	65:660	65:600	65:660	65:660
Final extension (ºC:s)	65:660	65:660	65:780	65:660	65:780	65:660	65:660	65:600	65:660	65:660
Hold	10ºC									

[1] Amplification of *CYP2B6* Frag1.
[2] Amplification of *CYP2B6* Frag3.
[3] Amplification of *CYP2B6* Frag4.
[4] Amplification of *CYP2B6* Frag5.
[5] Amplification of *CYP2B6* FragH1.
[6] Amplification of *CYP2B6* FragH2.
[7] Amplification of *CYP2A6* FragA.
[8] Amplification of *CYP2A6* FragD.
[9] Amplification of *CYP2A6*4 - CYP2A7*.
[10] Amplification of *CYP2A7-2A6* hybrids.

For *CYP2A6*, 9.2 kb-long amplicons (FragA), which comprise the *CYP2A6* gene as well as upstream and downstream non-coding regions, were generated following the long-range PCR (XL-PCR) protocols described by Wassenaar et al. (2016) with some modifications (Table 6.2). The XL-PCR forward and reverse primers used are shown in (Table 6.1). For samples where a gene duplication was predicted from the short read WGS, the duplicated *CYP2A6* gene copy was amplified with previously published primers (Fukami et al., 2007; Rao et al., 2000). See Figure S6.2 on page 179 for depictions of the *CYP2A6* XL-PCR fragments generated in this study.

The forward and reverse primers for both *CYP2B6* and *CYP2A6* were tailed with universal sequences on the 5' end to enable sample barcoding via a second PCR. The specific PCR cycling conditions are provided in Table 6.2. The components and concentrations of the PCR reaction mix are similar to those used for generating *CYP2D6* fragments (Chapter 5, subsubsection 5.3.5), except for the respective primer pairs.

6.3.4.2 Amplicon pooling, barcoding and sequencing

Amplicon pooling and barcoding was outsourced to Inqaba Biotech (Pretoria, South Africa) as the same company performed the subsequent HiFi sequencing. Quality control of the PCR

products from the first-round PCR was carried out using a 0.8% agarose gel for visual inspections and the Agilent 4200 TapeStation (Diagnostech, South Africa) for quantification following the D5000 Screen tape kit. For each participant, *CYP2B6* amplicons were pooled equimolarly with *CYP2A6* amplicons, and also with *CYP2D6* amplicons from the related subproject descibed in chapter 5. All amplicons were in the size range of 5–12 kb as the 13.5 kb-long *CYP2B6* Frag2 could not be amplified. The SMRT sequencing and CCS read generation followed the same steps described in Chapter 5, subsubsection 5.3.5.

6.3.4.3 HiFi sequence data analysis

Raw HiFi sequence data were demultiplexed and processed into individual samples according to the corresponding barcode sequences using the NGSutils next-generation sequencing data analysis software kit (Breese and Liu, 2013). *CYP2B6* and *CYP2A6* HiFi reads were aligned to the corresponding regions in GRCh38 and GRCh37, and also to the NG_007929.1 and NG_008377.1 reference sequences, respectively, using pbmm2 v1.7.0 (`https://github.com/PacificBiosciences/pbmm2`). Variant calling was done using DeepVariant (Poplin et al., 2018). Thereafter, variant phasing and read haplotagging was carried out for samples containing more than one heterozygous variant using WhatsHap (Patterson et al., 2015).

6.3.5 Variant functional prediction

CYP2B6 and *CYP2A6* variants were annotated using the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016). The functional effects of potential novel (not in PharmVar) non-synonymous variants were predicted using VEP plugins as in Chapter 5.

6.3.6 Statistical analysis

Allele frequencies were calculated and tested for deviation from Hardy Weinberg equilibrium (HWE) using the Fisher's exact test in R (`https://www.r-project.org`). A Pearson χ^2 analysis was used to determine significant differences in population *CYP2B6* and *CYP2A6* allele frequencies. P-values of < 0.05 were considered statistically significant for each test. For *CYP2B6*, allele frequency data for previous studies across SSA was obtained from the Pharmacogenomics Knowledgebase (PharmGKB) reference materials (`https://www.pharmgkb.org/page/cyp2b6RefMaterials`).

6.4 Results

6.4.1 Populations

The high coverage African genomes analysed in this study are representative of over 20 ethno-linguistic groups from 11 countries mainly across eastern, south/central and southern Africa as in chapter 5 (see Figure 5.1 on page 115 and Table 5.1 on page 116). Results from the *CYP2B6* and *CYP2A6* star allele analysis of high coverage genomes from African American/ Afro-Caribbean (ASW and ACB), European (EUR), admixed American (AMR), South Asian (SAS), and East Asian (EAS) individuals sequenced as part of the 1000 Genomes Project (Byrska-Bishop et al., 2021) were included for comparison.

6.4.2 *CYP2B6* star allele analysis and phenotype prediction

6.4.2.1 *CYP2B6* allele frequencies

Among the normal function *CYP2B6* star alleles, **1*, **2* and **17* were frequent in SSA (Table 6.3). However, the Berom in Nigeria and the Fon in Benin had no occurrence of **17*, while the participants from Cameroon lacked *CYP2B6*2* (Table 6.4), though the smaller sample size from Cameroon could have contributed to this observation. We found *CYP2B6*5* to be rare in SSA compared to European, admixed American and South Asian allele frequencies (Table 6.3).

The *CYP2B6*6* allele defined by rs3745274 (Q172H) and rs2279343 (K262R) was by far the most frequent decreased function *CYP2B6* allele in SSA participants (AF = 32.6%) and also across African American/Afro-Caribbeans (AF = 34.3%). The individuals of South Asian ancestry as well as the admixed Americans had comparable *CYP2B6*6* allele frequencies i.e. 36.3% and 34.7%, respectively. However, in comparison, the *CYP2B6*6* allele was present at significantly lower frequencies among the European individuals (AF = 22.2%, p = 4e-09) and East Asians (AF = 19.8%, p = 1.1e-13) genotyped in this study (Table 6.3). The *CYP2B6*6* allele frequency was non-uniform across SSA as it ranged from 21.7% in the Batswana to 47% among individuals from Burkina Faso (Table 6.4). Among other key decreased function alleles, we detected *CYP2B6*29* (hybrid deletion) in SSA (AF = 0.4%) and also among African Americans/Afro-Caribbeans (AF = 0.7%) and one admixed American individual. The *CYP2B6*9* allele which is defined by rs3745274 (Q172H) was present most frequently in SSA (AF = 3.2%) (Table 6.3).

Table 6.3: *CYP2B6* allele frequencies in sub-Saharan African populations compared to previous studies in this region and global populations.

		Allele frequencies in indicated populations (%)						
***CYP2B6* allele**	**CPIC function**	SSA in this study (n=935)	Previous SSA studies [a] (n=1860)	African American/ Afro-Caribbean (n=153)	European (n=499)	Admixed American (n=344)	East Asian (n=501)	South Asian (n=479)
**1*	Normal	42.8	31.5	42.8	53.5	47.5	66.6	42.9
**2*	Normal	4.2	3.1	3.3	5.3	2.6	4.2	4.4
**5*	Normal	0.5	2	2	11.3	6.7	0.2	8.5
**17*	Normal	2.5	1.3	2.6	0	0.3	0	0
**32*	Normal	0.1		0	0	0	0	0
**6*	Decreased	32.6	37	34.3	22.2	34.7	19.8	36.3
**7*	Decreased	0.2	1.6	0	0	0	0	0.1
**9*	Decreased	3.2		0.3	0.6	1.2	0.3	1
**19*	Decreased	0.2	1.6	0	0	0	0	0
**20*	Decreased	0.1	1.6	0	0	0.1	0	0
**26*	Decreased	0	0	0	0	0	0.5	0
**29*	Decreased	0.4	0	0.7	0	0.1	0	0
**36*	Decreased	0.8		1	0	0.1	0	0.1
**12*	No function	0	0	0	0	0.1	0	0
**13*	No function	0	0	0.7	0.1	0	0	0.1
**18*	No function	9.5	5.8	7.5	0	1	0	0
**24*	No function	0	0	0	0	0	0.1	0
**28*	No function	0	0.6	0	0	0	0	0
**4*	Increased	0.1	0	0.3	3	1	6.3	3.7
**22*	Increased	1.1	3.1	2	0.9	0.6	0.2	1.7
**3*	Uncertain	0	0	0.3	0.1	0	0	0.2
**10*	Uncertain	0	0	0	0.6	1.3	0.1	0.1
**11*	Uncertain	0.1	7.1	0	0.4	0	0	0
**15*	Uncertain	0	1.5	0	0.4	0.4	0	0
**21*	Uncertain	0	1.3	0	0	0	0	0
**23*	Unknown	0	0	0	0	0	0.2	0.1
**27*	Uncertain	0	0.6	0.3	0	0	0	0
**33*	Uncertain	0.1		0.3	0	0	0	0

Table 6.4: *CYP2B6* allele frequencies across all sub-Saharan African populations included in the study.

		Allele frequencies (%)												
***CYP2B6* allele**	**CPIC function**	Benin (n=50)	BRN (n=45)	Burkina Faso (n=33)	Ghana (n=26)	Cameroon (n=24)	Zambia (n=41)	Botswana (n=46)	South Africa (n=151)	ESN (n=97)	YRI (n=106)	GWD (n=111)	MSL (n=82)	LWK (n=95)
**1*	Normal	35	46.7	31.8	42.3	43.8	40.2	53.3	41.4	45.4	35.8	43.2	46.3	47.4
**2*	Normal	6	1.1	6.1	1.9	0	2.4	4.3	4.6	2.6	5.2	5	2.4	7.4
**5*	Normal	2	0	1.5	1.9	0	0	0	0.3	0	0.5	1.4	0.6	0
**17*	Normal	0	0	4.5	5.8	4.2	6.1	2.2	3	1.5	4.7	2.3	1.2	0.5
**32*	Normal	0	0	0	0	0	0	0	0.3	0	0	0	0	0
**6*	Decreased	42	37.8	47	38.5	27.1	28	21.7	32.5	39.7	34	26.1	28.7	31.1
**7*	Decreased	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.6	0
**9*	Decreased	0	4.4	3	3.8	0	4.9	1.1	2.6	1.5	4.7	4.5	4.3	3.2
**19*	Decreased	0	0	0	0	0	0	0	0	0	0.5	0.5	1.2	0
**20*	Decreased	0	0	0	0	0	1.2	0	0.3	0	0	0	0	0
**26*	Decreased	0	0	0	0	0	0	0	0	0	0	0	0	0
**29*	Decreased	0	2.2	0	0	0	1.2	0	0	0.5	0.9	0.5	0	0
**36*	Decreased	2	3.3	0	0	2.1	1.2	1.1	0	0	0	0.9	0	2.1
**12*	No function	0	0	0	0	0	0	0	0	0	0	0	0	0
**13*	No function	0	0	0	0	0	0	0	0	0	0	0	0	0
**18*	No function	11	2.2	3	5.8	18.8	14.6	13	12.9	5.2	11.8	9.9	10.4	6.3
**24*	No function	0	0	0	0	0	0	0	0	0	0	0	0	0
**28*	No function	0	0	0	0	0	0	0	0	0	0	0	0	0
**4*	Increased	0	0	0	0	0	0	0	0	0	0	0.5	0	0
**22*	Increased	0	0	1.5	0	2.1	0	1.1	0.7	2.1	0.9	1.8	2.4	0.5
**3*	Uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**10*	Uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**11*	Uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0.5
**15*	Uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**21*	Uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 6.4 – continued

CYP2B6 allele	CPIC function	Allele frequencies (%)												
		Benin (n=50)	BRN (n=45)	Burkina Faso (n=33)	Ghana (n=26)	Cameroon (n=24)	Zambia (n=41)	Botswana (n=46)	South Africa (n=151)	ESN (n=97)	YRI (n=106)	GWD (n=111)	MSL (n=82)	LWK (n=95)
**23*	Unknown	0	0	0	0	0	0	0	0	0	0	0	0	0
**27*	Uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**33*	Uncertain	0	0	0	0	2.1	0	0	0	0	0	0	0	0

BRN, Berom in Nigeria; ESN, Esan in Nigeria; GWD, Gambian in Western Division (Mandinka); LWK, Luhya in Webuye (Kenya); MSL, Mende in Sierra Leone; YRI, Yoruba in Nigeria.

*CYP2B6*18* which is defined by rs28399499 (I328T) was the only no function *CYP2B6* allele present in SSA (AF = 9.5%). The frequency of **18* was slightly lower in African Americans/Afro-Carribeans (AF = 7.5%) but this difference was not statistically significant (p = 0.3). In comparison, *CYP2B6*18* was found to be virtually absent among the Europeans, East Asians and South Asians represented in the 1000 Genomes Project dataset (Table 6.3). The frequency of **18* varied across the SSA populations, ranging from 2.2% among the Berom to 18.8% among the participants from Cameroon (Table 6.4).

As regards to the increased function *CYP2B6* alleles, *CYP2B6*4* — rs2279343 (K262R) without rs3745274 (Q172H) in LD — was found in only one GWD participant (see Table 6.4) across SSA. Conversely, *CYP2B6*22* (rs34223104) was more common and present in all SSA populations included in this study (combined AF = 1.1%) except for the Fon in Benin, Berom in Nigeria, Ghanaian and Zambian participants. In comparison, *CYP2B6*4* is more common among Europeans (AF = 3%), East Asians (AF = 6.3%) and South Asians (AF = 3.7%), while **22* was estimated to be relatively rarer across the global populations (Table 6.3), with the highest frequency (2%) being among the African Americans/Afro-Caribbeans.

Among the *CYP2B6* alleles with unknown/uncertain function currently catalogued by PharmVar, we only identified **11* and **33* in SSA, both of which were singletons. This was in contrast to the expected frequency for **11* (7.1%) inferred from previous studies across SSA (Table 6.3).

6.4.2.2 Predicted *CYP2B6* phenotypes

The distributions of predicted CYP2B6 metaboliser phenotypes relative to efavirenz response are depicted in Figure 6.2 and Figure 6.3. Two hundred seven participants (21.5%) were predicted to be CYP2B6 poor metabolisers (PMs) across the SSA populations in the study. This was comparable to the proportion of CYP2B6 PMs in the African Americans/Afro-Caribbeans but significantly higher than the proportion of PMs in European, admixed American, East Asian and South Asian populations (Figure 6.2). The proportion of predicted CYP2B6 PMs varied across SSA, with the lowest proportion being among the Batswana participants (12.8%), and the highest PM proportion observed among the participants from South Africa (25.8%), Benin (28%) and Burkina Faso (30.3%) (Figure 6.3. These differences were largely due to the variability in the distribution of *CYP2B6*6/*6*, **6/*18* and **18/*18* diplotypes.

For the CYP2B6 intermediate metaboliser (IM) phenotype, participants from across SSA had a considerably higher frequency of predictive diplotypes compared to Europeans and East Asians (Figure 6.2). The proportion of IMs in SSA populations was comparable to that in African Americans/Afro-Caribbeans, admixed Americans and South Asians (Figure 6.2). The distribution of IMs was diverse across SSA (Figure 6.3) — the relative proportion of IMs was particularly high among participants from Ghana (57%), Cameroon (50%), the Yoruba (52.8%), and the Esan

(52.5%), while the Berom (38.8%), Luhya (39.4%) and South African participants (41.3%) had a relatively lower proportion IMs.

The proportion of SSA participants (1.4%) with the CYP2B6 rapid metaboliser (RM) phenotype was significantly lower than that of Europeans (4.6%), South Asians (6.5%) and East Asians (8.7%) (Figure 6.2). All the SSA populations in this study had $\leq$ 3 participants with the CYP2B6 RM phenotype and in particular, it was not observed among the Fon in Benin, Berom in Nigeria, and participants from Ghana, Cameroon and Zambia (Figure 6.3). Of note, the ultrarapid metaboliser (UM) phenotype was only predicted in one participant across all SSA populations in the study. This is an Esan participant that had the *CYP2B6*22/*22* diplotype.

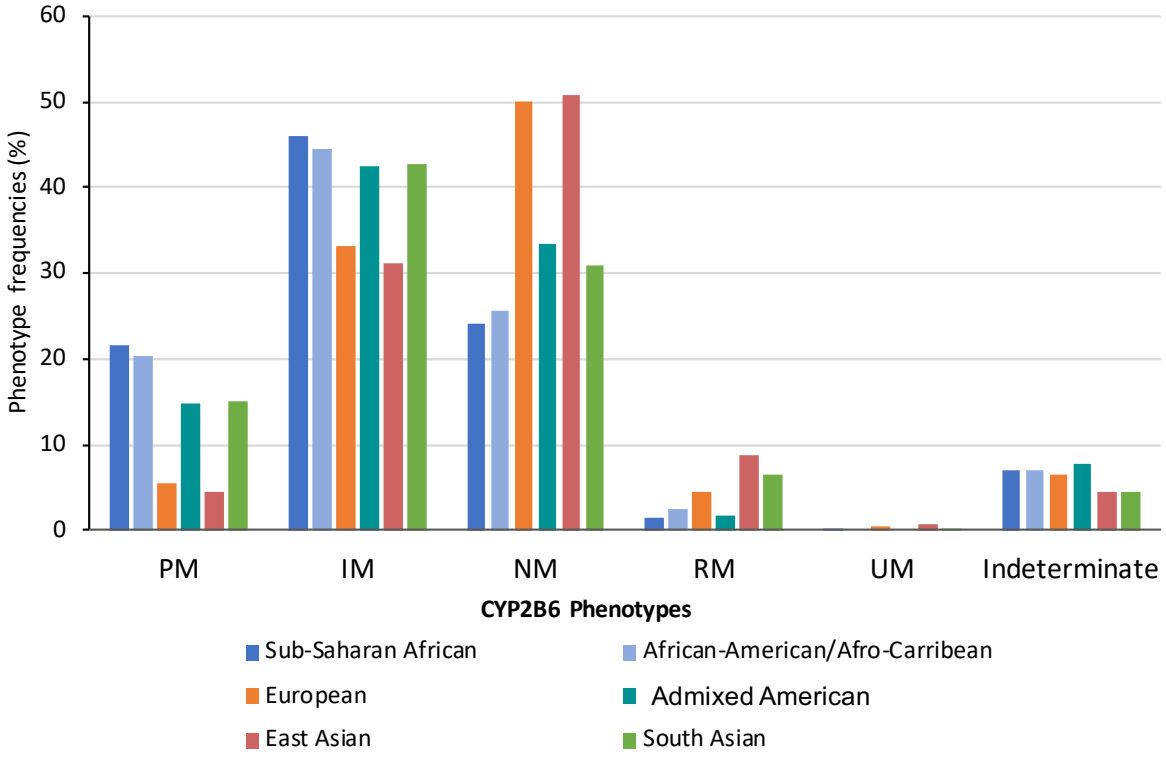


Figure 6.2: Distribution of *CYP2B6* phenotypes predicted from diplotypes called from high-coverage WGS data. As shown in the figure, SSA poplations have a higher proportion of CYP2B6 poor metabolisers and intermediate metabolisers compared to other biogeographical groups. This is in part accounted for by the high frequency of the decreased function *CYP2B6*6* across Africa. PM, poor metaboliser; IM, intermediate metaboliser; NM, normal metaboliser; RM, rapid metaboliser; UM, ultrarapid metaboliser.

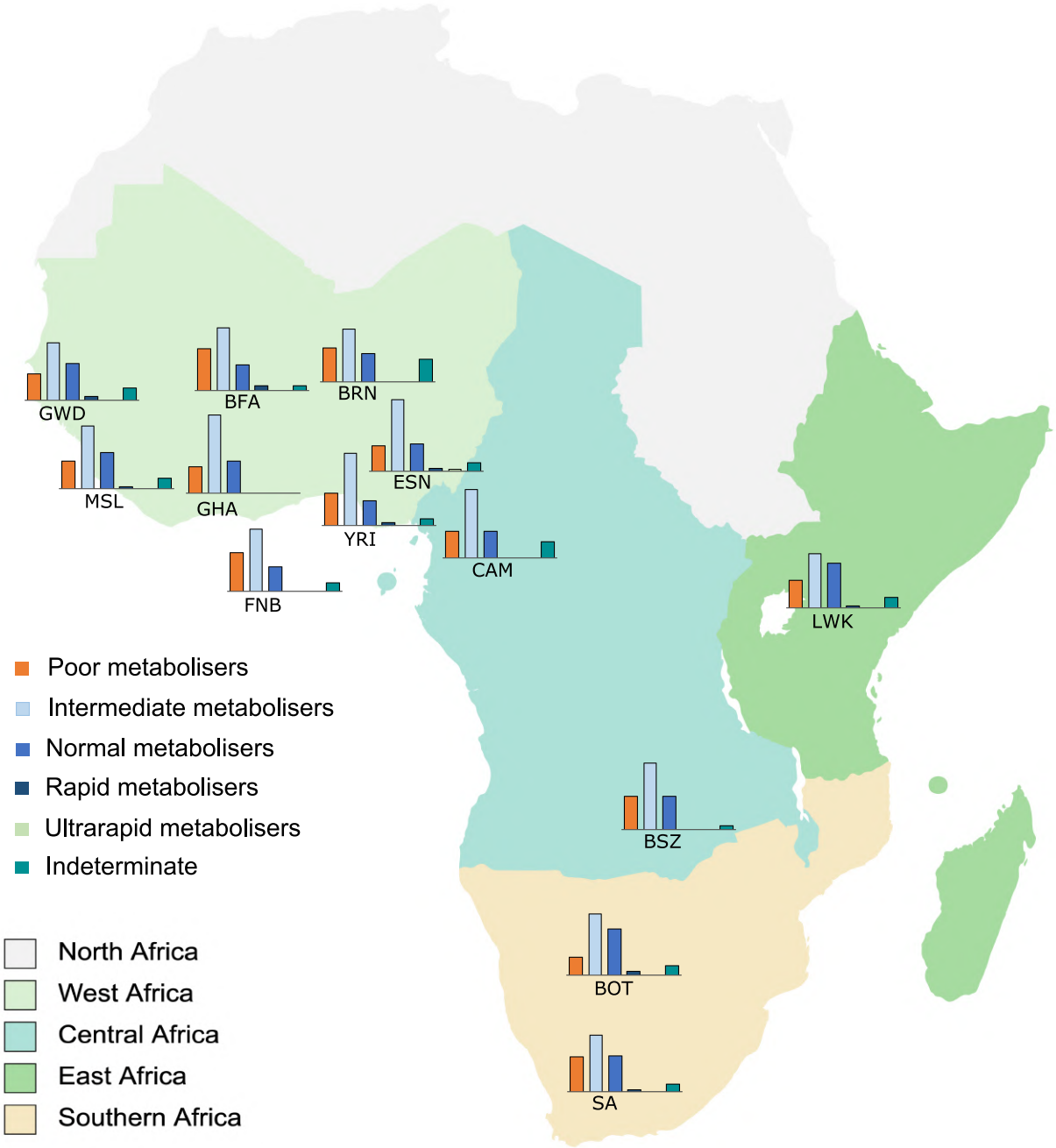


Figure 6.3: CYP2B6 phenotype distribution across the SSA populations in the study. BFA, participants from Burkina Faso; BOT, participants from Botswana; BRN, Berom in Nigeria; BSZ, Bantu-speakers in Zambia; CAM, Cameroonian participants; ESN, Esan in Nigeria; FNB, Fon in Benin; GHA, Ghanaian partcipants; GWD, Gambian in Western Division (Mandinka); LWK, Luhya in Webuye (Kenya); MSL, Mende in Sierra Leone; YRI, Yoruba in Nigeria.

6.4.2.3 Potential novel *CYP2B6* haplotypes

We identified 19 high-confidence potential novel *CYP2B6* haplotypes in SSA populations based on the high coverage Illumina WGS data used in the study. Fourteen of these alleles were fully phased computationally while 5 were observed in individuals whose diplotypes could not be resolved i.e. the background allele on which the novel variants occured could not be determined (Table 6.5). For the phased potential novel *CYP2B6* alleles, the phasing of novel core variants was made possible either by having homozygous background alleles in the same participant or observing multiple participants with the same novel core variant while also sharing one background allele. However, suballele definitions could not be ascertained by this approach. The core variants in question are not novel *per se* but are rather nonsynonymous variants that are either not currently catalogued as allele-defining by PharmVar or have been found in different allelic combinations in our study. The core variants of these potential novel *CYP2B6* alleles were predicted to be deleterious by at least one VEP plugin used in the study, except for rs33973337 (T26S), rs541486480 (K53Q), rs537265436 (F202L), rs1599849465 (E240D), and rs3211371 (R487C) (Table 6.6).

In addition to the aforementioned 19 potential novel alleles, we found a potential novel *CYP2B6-2B7* hybrid duplication in 21 individuals across various SSA populations. Using short-read data, we could not accurately estimate the breakpoints for this hybrid duplication in the genomes of these individuals or confirm whether it was similar to *CYP2B6*30* (Martis et al., 2013).

Table 6.5: Potential novel African-specific *CYP2B6* alleles inferred from Illumina WGS data

Haplotype	Background *CYP2B6* allele	Additional core variant(s)	Allele count	Country/Dataset
1	**1*	rs541486480∼5164A>C	2	Nigeria (1000G)
2	**1*	rs373926269∼17867T>C	1	Gambia (1000G)
3	**1*	rs370958436∼18040C>T	1	Nigeria
4	**1*	rs766630605∼20648G>A	1	Nigeria
5	**1*	rs537265436∼20726T>C	2	Botswana, South Africa
6	**1*	rs1599849465∼22995A>C	1	South Africa
7	**1*	rs150072531∼26413A>G	2	Gambia (1000G), Benin
8	**2*	rs3211371∼30512C>T	2	Burkina Faso, Cameroon
9	**6*	rs572134005∼17785G>A	1	Sierra Leone
10	**6*	rs142421637∼17856C>T	5	Botswana, South Africa, Kenya (1000G, SGDP)
11	**6*	rs58871670∼20669G>A	4	Gambia (1000G), Nigeria (1000G)
12	**22*	rs141666881∼18137G>A	6	Gambia (1000G, SGDP), Nigeria (1000G), Sierra Leone (1000G)
13	**22*	rs34698757∼23790C>G	1	Sierra Leone (1000G)
14	**22*	rs147991149∼30380C>T	2	South Africa

Unresolved *CYP2B6* sub-Saharan African diplotypes with potential novel alleles

#	Background *CYP2B6* alleles	Additional core variant(s)	Number of participants	Country/dataset
1	**1/*18*	rs34698757∼23790C>G	1	Botswana
2	**1/*6*	rs183427203∼17823C>T	1	Kenya (1000G)
3	**1/*6*	rs33973337∼5083A>T	1	Cameroon
4	**2/*9*	rs28399499∼26018T>C	1	Botswana/Namibia (SGDP)
5	**6/*17*	rs45459594∼20668C>G + rs36079186∼20715T>C	1	Kenya (1000G)
6	**1 or *22*	*CYP2B6/2B7* hybrid (Various breakpoints)	21	Various

Coriell IDs for samples that could be used as potential GeT-RM reference materials for *CYP2B6* genotyping are provided in Table S6.1.

Table 6.6: VEP plugin predictions of the deleteriousness of core variants defining potential novel African-specific *CYP2B6* alleles identified in the study.

Core_variants	Consequence	CADD	SIFT	PolyPhen-2	LRT	PROVEAN	VEST3
rs33973337	missense (T26S)						
rs541486480	missense (K53Q)						
rs572134005	missense (R85Q)	X	X				
rs183427203	missense (R98W)	X	X	X	X	X	X
rs142421637	missense (R109W)	X	X	X			
rs373926269	splice donor	X					
rs370958436	stop gained	X			X		X
rs141666881	missense (R158Q)						
rs766630605	missense (A176T)			X	X		
rs45459594	missense (I182M)		X	X	X		
rs58871670	missense (V183I)				X		
rs36079186	missense (M198T)	X	X				
rs537265436	missense (F202L)						
rs1599849465	missense (E240D)						
rs34698757	missense (T306S)	X	X	X	X	X	X
rs28399499	missense (I328T)	X	X	X	X	X	X
rs150072531	missense (H397R)					X	
rs147991149	missense (R443C)	X	X	X	X	X	
rs3211371	missense (R487C)						

X, indicates that the variant was predicted to be deleterious by the corresponding plugin; The deleterious score thresholds used are the same as those in chapter 5.

6.4.3 Long-read-based characterisation of novel *CYP2B6* alleles

As the optimisation for an CYP2B6 Frag2 (overlapping exon 1 to exons 2-3) was not successful, it was not possible to fully characterise haplotypes with multiple novel core variants in exon 1 and other exons. The targeted SMRT sequencing enabled further characterisation of Haplotype 5 and Haplotype 8 identified in participants from South Africa and Burkina Faso, respectively (Table 6.5 and Figure 6.4). The rs537265436 (NG_007929.1: g.20726T>C, F202L) variant which defines Haplotype 5 (Figure 6.4) was replicated in the targeted HiFi data of a South African participant with *CYP2B6*1/*1* background diplotype. Subvariants in the same phase set as rs537265436 were ascertained from WhatsHap phased output and are summarised in Figure 6.4. For Haplotype 8 (Figure 6.5), our HiFi data confirmed that rs3211371 (NG_007929.1: g.30512C>T, R487C) was in phase with the *CYP2B6*2*–defining variant, rs8192709 (NG_007929.1: g.5071C>T, R22C). For the other participants that have been successfully resequenced so far, the star alleles predicted from the short-read WGS data were concordant with the ones identified from the long read data. Moreover, we inferred more phasing information for suballeles from the HiFi data mainly from exons 2-9.

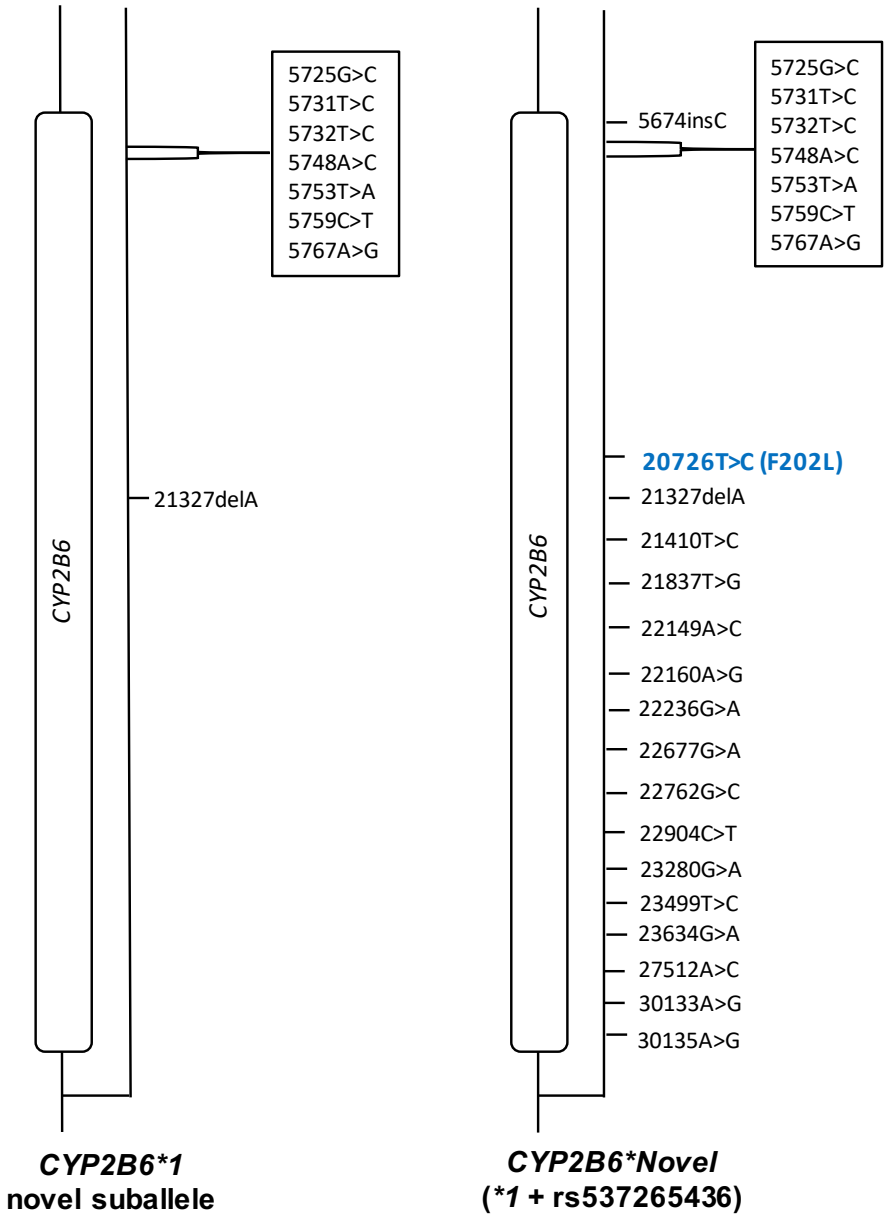


Figure 6.4: Characterisation of a *CYP2B6* novel allele involving rs537265436 via targeted SMRT sequencing. The defining rs537265436 (F202L) variant was identified on a *CYP2B6*1* background in a South African participant with a novel *CYP2B6*1* suballele on the other arm of chromosome 19.

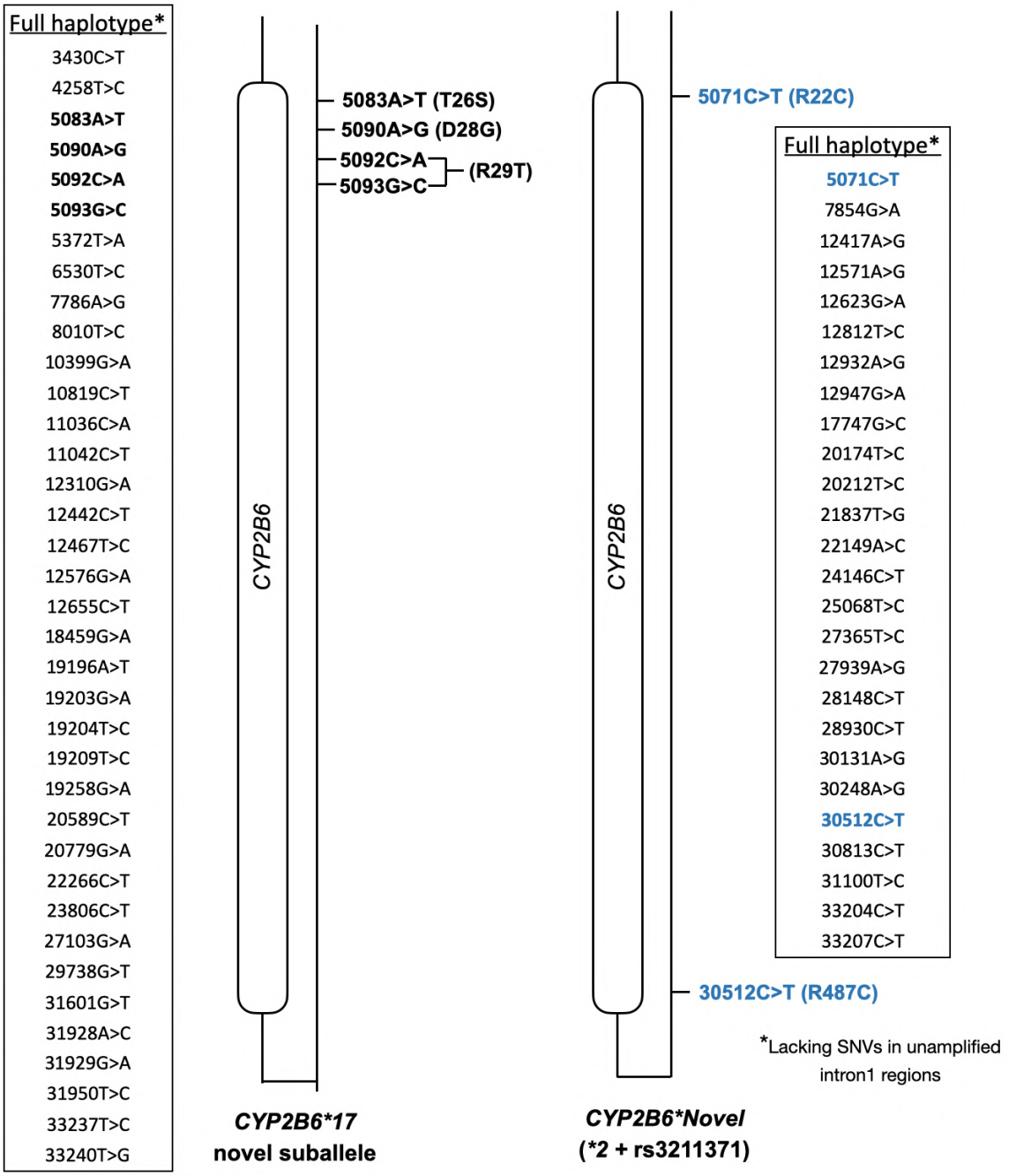


Figure 6.5: Characterisation of a *CYP2B6* novel allele involving rs3211371 via targeted SMRT sequencing. The defining rs3211371 (R487C) variant was confirmed to be in phase with the *CYP2B6**2*-defining variant rs8192709 (R22C) after targeted resequencing of a participant from Burkina Faso. The second *CYP2B6* haplotype in the participant was characterised as a novel *CYP2B6*17* suballele as shown in the figure.

6.4.4 *CYP2A6* star allele analysis

6.4.4.1 *CYP2A6* allele frequencies

The frequencies of established *CYP2A6* star alleles are summarised in Table 6.7 for SSA and other global populations (for comparison) represented in this study. *CYP2A6*1B* (58bp gene conversion in 3' UTR), which causes increased *CYP2A6* mRNA stability, was observed at a frequency of 6.1% in SSA. This allele was observed at significantly higher frequencies in European, East Asian, and South Asian populations (Table 6.7). The frequency of *CYP2A6*1B* varied across SSA populations in the study, ranging from 1.5% among the LWK to 11% among the Fon in Benin (Table 6.8). Among the established duplication alleles, we called *CYP2A6*1x2* (causes increased mRNA expression) but the algorithms used did not separate *CYP2A6*1x2A* from *CYP2A6*1x2B*. The *CYP2A6*1x2* allele was present at a frequency of 0.8% in SSA which was significantly less than the frequency in ACB and ASW (p = 0.01), comparable to the frequency in Europeans, admixed Americans and South Asians in this study, but absent from the East Asian populations. Among the SSA participants, *CYP2A6*1x2* had the highest frequency among the participants from Ghana (AF = 1.9%) and the MSL (AF = 1.8%) but it was not observed among the Fon in Benin, Bantu speakers in Zambia and the Cameroonian participants (Table 6.8).

Among the extensive amount of *CYP2A6* alleles that cause decreased CYP2A6 expression and/or activity, **9* which is defined by the rs28399433 variant in the TATA box and **17* (rs28399454, V365M) were the most frequent across SSA (Table 6.7 and Table 6.8). The frequency of the *CYP2A6*9* haplotype in SSA (7.5%) was comparable to that in other biogeographical groups, except in the East Asian populations (AF = 19%). In contrast, *CYP2A6*17* which is known to be African-specific was observed at a frequency of 11% in SSA, 12.3% in African Americans/Afro-Caribbeans and absent among the Europeans, East Asians and South Asians (Table 6.7). Across SSA, the *CYP2A6*9* frequency was lower among the Berom, Batswana and participants from Burkina Faso compared to other populations in the study (Table 6.8).

We observed the *CYP2A6*4* (*CYP2A6* gene deletion) at a frequency of 3.1% in SSA, which was comparable to the **4* frequency in the ACB and ASW, and South Asian individuals. However, *CYP2A6*4* was less frequent among the European populations and highest among the East Asian populations (Table 6.7). Among the SSA participants in this study, *CYP2A6*4* was most frequent among the Berom and the Bantu-speakers in Zambia (AF = 6.1%), but it was not observed among the participants from Ghana in this study.

Other largely African-specific *CYP2A6* star alleles that were observed in SSA in our study include **20*, **24*, **25*, **26*, **27*, **28*, **31*, **35*, **39* and **41*.

Table 6.7: *CYP2A6* allele frequencies in sub-Saharan African populations as compared to global populations.

		Allele frequencies (%)					
***CYP2A6* allele**	**Functional impact on CYP2A6**	**SSA (n=957)**	**African American/ Afro-Caribbean (n=155)**	**European (n=498)**	**Admixed American (n=344)**	**East Asian (n=503)**	**South Asian (n=486)**
**1*	Normal enzyme activity	55.7	54.8	51.3	38.8	32	46.5
**1B*	Increased mRNA stability	6.1	8.4	27.7	41.7	21.6	26.9
**1x2*	Increased mRNA expression	0.8	2.6	0.9	0.7	0	0.5
**2*	Substantially decreased enzyme activity	0	0.3	3.2	0.7	0	0.6
**4*	no mRNA expression	3.1	2.3	0.9	1.5	11.8	3.8
**5*	Decreased enzyme activity	0	0	0	0	0.1	0.1
**7*	Decreased enzyme activity	0	0	0.2	0	8	0.2
**8*	Normal enzyme activity	0	0	0	0	0.1	0
**9*	Decreased mRNA expression	7.5	7.7	7	9.4	19	12.9
**10*	Inactive enzyme	0	0	0	0	3	0.2
**11*	Decreased enzyme activity	0	0	0	0	0.8	0
**12*	Decreased enzyme activity	0.1	0	2	1.9	0	0.9
**13*	Unknown/ uncertain	0	0	0	0	0.2	0
**14*	Unknown/ uncertain	0	1	3.2	1.6	0	2.3
**15*	Unknown/ uncertain	0	0	0	0	1.1	0
**17*	Substantially decreased enzyme activity	11	12.3	0	0.4	0	0
**18*	Decreased enzyme activity	0.3	0	1.4	1.3	0.6	1.7
**19*	Decreased enzyme activity	0	0	0	0	0.4	0
**20*	Substantially decreased protein levels	0.8	0.3	0	0	0	0
**21*	Decreased enzyme activity	0	0	1.2	0.6	0	0.8
**23*	Decreased enzyme activity	1.3	3.2	0	0	0	0
**24*	Decreased enzyme activity	1	0.6	0	0.1	0	0
**25*	Decreased enzyme activity	0.3	0.3	0	0	0	0
**26*	Decreased enzyme activity	0.3	0.6	0	0	0	0
**27*	Decreased enzyme activity	0.9	0.6	0	0	0	0
**28*	Decreased enzyme activity	2.4	0.6	0	0.3	0	0.1
**31*	Unknown/ uncertain	2.3	1.3	0	0	0	0
**34*	Unknown/ uncertain	0.1	0.3	0	0.1	0	0.1
**35*	Decreased enzyme activity	2.8	1.6	0.8	0.3	0.2	0
**36*	Unknown/ uncertain	0	0	0	0	0.5	0
**37*	Unknown/ uncertain	0	0	0	0	0.2	0
**39*	Decreased enzyme activity	0.1	0	0	0	0	0
**41*	Decreased enzyme activity	0.1	0	0	0	0	0

Table 6.8: *CYP2A6* allele frequencies across all sub-Saharan African populations included in the study.

***CYP2A6* allele**	**Functional impact**	**Allele frequencies (%)**												
		Benin (n=50)	BRN (n=49)	Burkina Faso (n=33)	Ghana (n=26)	Cameroon (n=26)	Zambia (n=41)	Botswana (n=47)	South Africa (n=155)	ESN (n=98)	YRI (n=108)	GWD (n=113)	MSL (n=84)	LWK (n=98)
**1*	Normal enzyme activity	53	54.1	56.1	59.6	53.8	56.1	59.6	60.3	54.1	56.9	52.2	51.8	62.8
**1B*	Increased mRNA stability	11	7.1	7.6	3.8	9.6	8.5	4.3	6.5	5.6	7.4	3.1	5.4	1.5
**1x2*	Increased mRNA expression	0	1	1.5	1.9	0	0	1.1	1	1	0.5	0.4	1.8	0.5
**2*	Substantially decreased enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0	0
**4*	no mRNA expression	2	6.1	1.5	0	1.9	6.1	5.3	1	3.1	3.7	2.7	5.4	3.1
**5*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0	0
**7*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0	0
**8*	Normal enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0	0
**9*	Decreased mRNA expression	10	4.1	3	9.6	9.6	9.8	5.3	8.4	5.1	7.9	8.8	8.3	8.2
**10*	Inactive enzyme	0	0	0	0	0	0	0	0	0	0	0	0	0
**11*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0	0
**12*	Decreased enzyme activity	0	0	0	0	0	0	1.1	0	0	0	0	0	0
**13*	Unknown/ uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**14*	Unknown/ uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**15*	Unknown/ uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**17*	Substantially decreased enzyme activity	13	10.2	19.7	11.5	17.3	6.1	6.4	8.7	14.8	7.9	12.4	12.5	9.7
**17x2*	Unknown/ uncertain	1	0	0	0	0	0	0	0	0	0	0	0	0
**18*	Decreased enzyme activity	0	1	0	0	0	0	0	0	0.5	0.5	0.4	0	0.5
**19*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0	0
**20*	Substantially decreased protein levels	0	0	0	0	0	0	0	0.3	1.5	0.5	2.7	1.2	1
**21*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0	0
**23*	Decreased enzyme activity	4	0	1.5	1.9	0	1.2	0	0	1.5	0.9	1.8	1.8	2
**24*	Decreased enzyme activity	2	4.1	1.5	0	0	0	1.1	0.3	0	1.4	0.9	1.8	1

Table 6.8 – continued

CYP2A6 allele	Functional impact	Allele frequencies (%)												
		Benin (n=50)	BRN (n=49)	Burkina Faso (n=33)	Ghana (n=26)	Cameroon (n=26)	Zambia (n=41)	Botswana (n=47)	South Africa (n=155)	ESN (n=98)	YRI (n=108)	GWD (n=113)	MSL (n=84)	LWK (n=98)
**25*	Decreased enzyme activity	0	1	0	0	0	0	0	0	0.5	0.5	0.4	1.2	0
**26*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0.5	1.3	0	0.5
**27*	Decreased enzyme activity	1	1	0	0	0	4.9	2.1	1.3	0	0	0.9	1.2	0.5
**28*	Decreased enzyme activity	0	6.1	1.5	3.8	0	3.7	1.1	0.3	1	4.2	4.4	1.8	3.1
**28x2*	Unknown/ uncertain	0	0	0	0	0	0	0	0.3	0	0	0	0	0
**31*	Unknown/ uncertain	2	1	1.5	1.9	3.8	2.4	5.3	5.5	0.5	1.4	1.3	1.2	0
**34*	Unknown/ uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0.5
**35*	Decreased enzyme activity	1	1	4.5	3.8	1.9	0	3.2	2.3	6.1	1.9	3.5	1.8	3.6
**36*	Unknown/ uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**37*	Unknown/ uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0
**39*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0.6	0
**41*	Decreased enzyme activity	0	0	0	1.9	0	0	0	0	0	0	0	0	0

BRN, Berom in Nigeria; ESN, Esan in Nigeria; GWD, Gambian in Western Division (Mandinka); LWK, Luhya in Webuye (Kenya); MSL, Mende in Sierra Leone; YRI, Yoruba in Nigeria.

6.4.4.2 Potential novel *CYP2A6* haplotypes

We identified 31 potential novel *CYP2A6* haplotypes in SSA populations in the study based on the high coverage short-read WGS data (Table 6.9 and Table S6.2 — page Table S6.2). We determined the phase for 28 of these alleles computationally based on the same strategy used for inferring the aformentioned potential novel *CYP2B6* alleles. The other 3 potential novel alleles occurred in unresolved diplotypes (Table 6.9).

Haplotype 13 (Table 6.9) was found in 13 participants in multiple SSA populations. This potential novel haplotype comprised the missense rs114558780 (NG_008377.1: g.10126C>T, T378I) variant on a *CYP2A6*1* background. Another potential novel haplotype that stood out, was Haplotype 16 (Table 6.9) with rs113558392 (NG_008377.1: g.9394T>G, V292G) on a *CYP2A6*1* background. This haplotype was found in 2 southern African individuals. In comparison, a potential duplication of this haplotype was found to be more common i.e. in 10 participants (see Table 6.9, Haplotype 17). Furthermore, we found potential duplications of *CYP2A6*17* and *CYP2A6*28* in this study (Table 6.9).

From the *in silico* 'novel' core variant effect predictions, rs58571639 (R311C), rs528089983 (T309I), rs554865113 (R129P), and rs558145012 (G36V) were predicted to be deleterious by all the VEP plugins used in this study, while rs143731390 (N438Y), rs8192730 (E419D), rs1809810 (N418D), rs2644906 (V292M), rs138978736 (Q239K), and rs72549435 (V110L) were predicted to be benign by all the tools (Table 6.10).

Table 6.9: Potential novel African-specific *CYP2A6* alleles inferred from Illumina WGS data

Haplotype #	*Background CYP2A6 allele*	Additional core variant(s)	Allele count	Country/dataset
1	**1*	rs558145012∼5128G>T	2	South Africa
2	**1*	rs72549435∼5615G>C	1	Botswana
3	**1*	rs780290198∼6720C>T	1	Cameroon
4	**1*	rs554865113∼6727G>C	1	Nigeria (1000G)
5	**1*	rs2545783∼6798G>T	2	South Africa, Nigeria (1000G)
6	**1*	rs557976670∼8433C>T	1	Kenya (1000G)
7	**1*	rs138978736∼8457C>A	2	Gambia (1000G), Nigeria (1000G)
8	**1*	rs58720852∼8515C>A	1	Sierra Leone (1000G)
9	**1*	rs111869995∼8567G>T	2	Sierra Leone (1000G), Zambia
10	**1*	rs528089983∼9445C>T	2	Nigeria (1000G)
11	**1*	rs58571639∼9450C>T	1	Kenya (1000G)
12	**1*	rs145036049∼10108G>A	1	Nigeria (1000G)
13	**1*	rs114558780∼10126C>T	13	Botswana, South Africa, Nigeria (1000G) Congo (SGDP), Gambia (1000G)
14	**1*	rs28399463∼10766A>G	4	Nigeria (1000G)
15	**1*	rs8192730∼10771G>C	1	Nigeria (1000G)
16	**1*	rs113558392∼9394T>G	2	Namibia (SGDP), South Africa
17	**1*	[(*1 + rs113558392∼9394T>G) x2]	10	Botswana, South Africa
18	**1*	rs61605570∼9990A>T	1	Nigeria (1000G)
19	**1*	rs533216061∼8508C>A + rs771986786∼7227C>G	3	South Africa
20	**1*	[(*1 + rs533216061∼8508C>A + rs771986786∼7227C>G) x 2]	1	South Africa
21	**5*	rs143731390∼11479A>T + rs72549435∼5615G>C]	1	Nigeria (1000G)
22	**9*	rs145308399∼5576G>A]	2	Nigeria (1000G), Sierra Leone (1000G)
23	**9*	rs1809810∼10689A>T	1	Nigeria
24	**17*	rs554920226∼10778C>A	1	Gambia (1000G)
25	**17*	Duplication of *17	1	Benin
26	**18*	rs4997557∼9400C>G + rs2644906∼9393G>A	3	Gambia (1000G), Nigeria (1000G), South Sudan (SGDP)
27	**28*	rs28399454∼10086G>A	1	Nigeria
28	**28*	Duplication of *28	1	South Africa

Unresolved *CYP2A6* sub-Saharan African diplotypes with potential novel alleles

#	*Background CYP2A6 allele*	Additional core variant(s)	Number of participants	Country/dataset
1	**9/*35*	rs137904044∼9983G>T	1	Nigeria (1000G)
2	**17/*35*	rs143067113∼5546G>A	1	Kenya (1000G)
3	**2/*27*	rs143731390∼11479A>T	1	Sierra Leone (1000G)
4	**9/*17*	Unresolved *CYP2A6* duplication	1	South Sudan (SGDP)

Table 6.10: VEP plugin predictions of the deleteriousness of core variants defining potential novel African-specific *CYP2A6* alleles identified in the study.

Core_variants	Consequence	CADD	SIFT	PolyPhen-2	LRT	PROVEAN	VEST4
rs143731390	missense (N438Y)						
rs554920226	missense (Q422K)				X		
rs8192730	missense (E419D)						
rs28399463	missense (N418D)				X		
rs1809810	missense (Y392F)						
rs114558780	missense (T378I)	X		X	X	X	
rs145036049	missense (R372H)				X		
rs28399454	missense (V365M)				X		
rs61605570	stop gained (R333*)	X					X
rs137904044	missense (E330D)	X	X		X		X
rs58571639	missense (R311C)	X	X	X	X	X	X
rs528089983	missense (T309I)	X	X	X	X	X	X
rs4997557	missense (T294S)				X		
rs113558392	missense (V292G)	X	X	X		X	
rs2644906	missense (V292M)						
rs111869995	missense (M275I)				X	X	
rs58720852	missense (T258K)	X	X		X	X	
rs533216061	missense (Q256K)				X		
rs138978736	missense (Q239K)						
rs557976670	missense (P231S)	X	X	X	X	X	
rs771986786	missense (Q218E)	X			X		X
rs2545783	missense (A153S)				X		
rs554865113	missense (R129P)	X	X	X	X	X	X
rs780290198	missense (L127F)	X	X	X	X		
rs72549435	missense (V110L)						
rs145308399	missense (E97K)	X		X	X		X
rs143067113	missense (V87I)				X		
rs558145012	missense (G36V)	X	X	X	X	X	X

X, indicates that the variant was predicted to be deleterious by the corresponding plugin; The deleterious score thresholds used are the same as those in Chapter 5.

6.4.5 Long-read-based characterisation of novel *CYP2A6* alleles

From the targeted SMRT sequencing of the *CYP2A6* XL-PCR fragments, we further characterised Haplotype 1 in Table 6.9 (see Figure 6.6) identified in two South African participants. This haplotype has rs558145012 (NG_008377.1: g.5128G>T, G36V) on a **1B* background. The subvariants in phase with this novel allele in one of the South African participants are shown in Figure 6.6. This participant had a novel *CYP2A6*17* suballele as the second haplotype (Figure 6.5).

The second novel major *CYP2A6* allele validated in this study is Haplotype 13 (Table 6.6) defined by rs114558780 (NG_008377.1: g.10126C>T, T378I) on a **1* background. The suballele depicted in Figure 6.6 is from a South African participant. However, this haplotype (allele count = 13) was also identified in participants from Botswana, Nigeria, Congo and the Gambia (Table 6.9), which adds another layer of evidence in terms of its existence and definition.

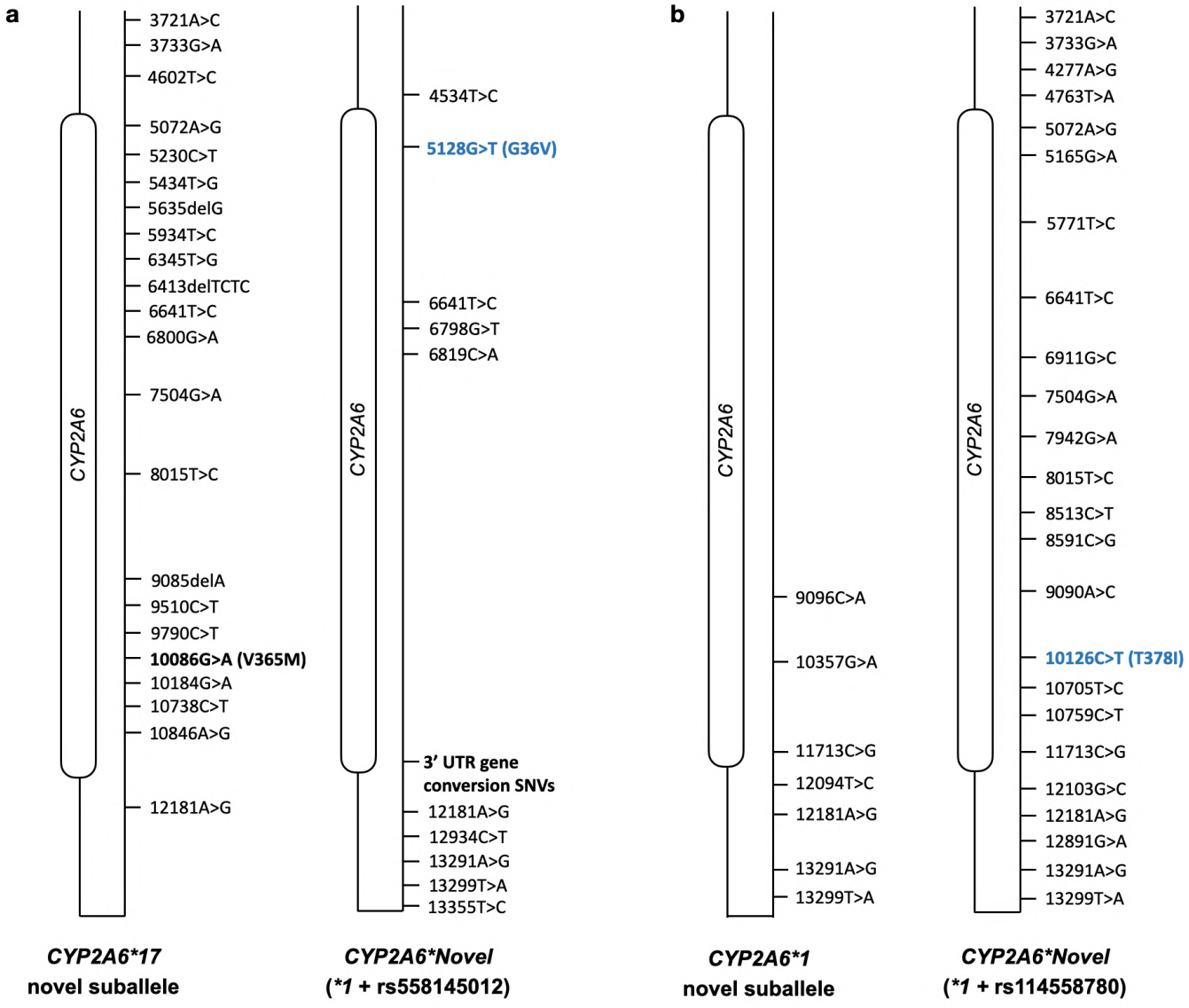


Figure 6.6: Characterisation of *CYP2A6* novel alleles via targeted SMRT sequencing. Panel (**a**) depicts the *CYP2A6* diplotype in a South African individual with the novel rs558145012 (G36V) core variant on a *CYP2A6*1B* background on one of the haplotypes. Having HiFi data facilitated the unambiguous read alignment spanning the entire *CYP2A6* region, including the 3'-UTR region despite the presence of the *CYP2A7* gene conversion. The second haplotype was characterised as a *CYP2A6*17* novel suballele. Panel (**b**) depicts the *CYP2A6* diplotype in a South African individual with the novel rs114558780 (T378I) core variant on a *CYP2A6*1* background on one of the haplotypes.

6.5 Discussion

CYP2B6 and *CYP2A6* are widely studied in pharmacogenomics as genetic variation in these genes is known to impact the metabolism and response to medications such as the antiretroviral efavirenz, and nicotine which is the major psychoactive component in cigarette smoke. In this study, we report the prevalence of *CYP2B6* and *CYP2A6* star alleles across SSA based on the analysis of 961 high coverage genomes representative of ethnically diverse populations from central, eastern, western, and southern Africa. These (short-read) data mainly include genomes generated by H3Africa projects (Choudhury et al., 2020; Ramsay et al., 2016) and other collaborations within Africa, and data from the 1000 Genomes Project Consortium (Byrska-Bishop et al., 2021). For *CYP2B6*, we further present the distribution of the efavirenz-based predicted phenotype distribution. Our analysis also includes comparisons between the *CYP2B6* and *CYP2A6* allele distributions in SSA and other global populations. In addition, we infer 50 potential novel African-specific alleles for *CYP2B6* and *CYP2A6* combined, and perform long-read-based characterisation for some of these alleles.

The known *CYP2B6* allele distributions across SSA populations are mainly from studies among participants from Ghana (K. Klein et al., 2005), Uganda (Mukonzo et al., 2014), Zimbabwe (Maimbo et al., 2012), Bantu-speakers in South Africa (Swart et al., 2013), and the five SSA populations represented in the 1000 Genomes Project Phase 3 dataset which largely comprises low coverage WGS data and whole exome sequence data (Auton et al., 2015). For *CYP2A6*, the available frequency data for individuals of African ancestry is predominantly from African American populations as reviewed by Tanner and Tyndale (2017). In comparison, this study presents *CYP2B6* and *CYP2A6* allele distributions from a more diverse set of genomes from resident SSA populations, including previously understudied populations e.g. the Berom in Nigeria, Fon in Benin, Bantu-speakers in Zambia, Batswana, participants from Cameroon, and South African populations represented in the AWI-Gen and CBRL datasets.

Our findings highlight a largely non-uniform allele distribution for both *CYP2B6* (Table 6.4) and *CYP2A6* (Table 6.8) across the SSA populations in this study. Some significant allele frequency differences are also observed between populations in neighbouring African countries and/or ethnolinguistic groups within the same country which typifies the complex pharmacogenomic variation landscape observed across Africa by previous studies (da Rocha, Othman, Botha, et al., 2021; Rajman et al., 2017).

As expected *CYP2B6*6* was the most frequent decreased function allele in SSA while the no function *CYP2B6*18* allele was also common but with varying frequencies. These two alleles are partly responsible for the high proportion of efavirenz poor metaboliser phenotypes predicted in SSA in this study (Figure 6.2) which is consistent with other research studies in SSA (Dhoro

et al., 2015; Maimbo et al., 2012; Nyakutira et al., 2008; Peko et al., 2019; Röhrich et al., 2016). However, interethnic differences in the frequency of *CYP2B6*6* (Table 6.4) and presence of potential novel alleles on a *CYP2B6*6* background (Table 6.4) exemplify the caveats of blanket precision medicine implementation strategies. Most commercial pharmacogenetic testing panels only include *CYP2B6*6* and **18*. However, based on the results in this study, using these existing panels among SSA populations, where *CYP2B6*9*, **20*, **22*, **29*, and **36* frequencies can be >1%, could lead to inaccurate star allele calls.

For *CYP2A6*, the **1B* and **1x2* alleles which are associated with greater *in vivo* nicotine metabolism (Tanner and Tyndale, 2017) were both detected from the WGS data used in this study. *CYP2A6*1B* (58bp gene conversion in 3'-UTR) was challenging to call as it occurs in LD with multiple other alleles, and also causes read misalignment with *CYP2B7*. The *CYP2B6*1B* 3'-UTR conversion is associated with increased mRNA stability, thus contributing to increased CYP2A6 expression (J. Wang et al., 2006). In general, **1B* occurred at a lower frequency among SSA populations in comparison to other global populations. Frequencies for *CYP2A6*1B* and the rarer *CYP2A6*1x2* varied considerably across the SSA populations in this study. The allele calling tools we used did not differentiate *CYP2A6*1x2A* from *CYP2A6*1x2B* both of which result from unequal crossover of *CYP2A6* and the neighbouring *CYP2A7* pseudogene during recombination. The reciprocal of this unequal cross over is the no function *CYP2A6*4* (gene deletion) allele, which we observed at frequencies as high as 6.1% in the Berom, and BSZ, and contrasting lower frequencies among some of the other SSA populations. For the aforementioned *CYP2A6* alleles and most of the alleles associated with decreased CYP2A6 activity, the frequencies available to date for individuals of African ancestry have mostly been representative of African Americans/Afro-Caribbeans (Tanner and Tyndale, 2017). Our work therefore provides insights into the distribution of these alleles in continental African populations.

We observed unique distributions of the CYP2B6 metaboliser phenotypes across different SSA populations (Figure 6.3) and also in comparison to other biogeographical groups (Figure 6.2). This was mainly due to the differences in the diplotype frequencies and largely supported the estimates in the PharmGKB and CPIC CYP2B6 reference materials (`https://www.pharmgkb.org/page/cyp2b6RefMaterials`). The high proportion of participants with CYP2B6 poor metaboliser phenotypes in SSA in relation to efavirenz response emphasises the need for catalysing precision medicine implementation across Africa through more sequencing studies especially for understudied ethnolinguistic groups. The predicted CYP2B6 phenotypes should be interpreted with caution as there are a number of non-genetic factors not investigated in this study that could influence the CYP2B6 phenotype (Croom et al., 2009; Desta et al., 2019; Thomford et al., 2016).

This study inferred multiple novel African-specific alleles for both *CYP2B6* and *CYP2A6*. The

core variants defining these alleles were all rare and are not novel per se, but are nonsynonymous variants that are either not currently catalogued as allele-defining by PharmVar or have been found in different combinations in our study. To our knowledge, this is the first study to perform targeted SMRT sequencing to characterise novel *CYP2A6* haplotypes. Two of the novel major *CYP2A6* alleles inferred from the short-read data were further characterised via SMRT sequencing, and multiple novel suballeles were ascertained. *CYP2B6* targeted SMRT sequencing presented more challenges detailed below in the limitations but it also provided resolution for two novel *CYP2B6* alleles. The process of submitting these novel alleles to PharmVar is ongoing. The presence of these rare previously uncharacterised alleles exemplify the pharmacogene allelic diversity observed among Africans by several studies (da Rocha, Othman, Botha, et al., 2021; Matimba et al., 2009; Rajman et al., 2017), and can impact the response to medications metabolised by CYP2B6 and CYP2A6.

There were several limitations in this study. Firstly, our analysis was focused on populations within SSA and none from Northern Africa. Despite the multiple sources of the genomic data we used, we had very few participants from Nilo-Saharan, Afroasiatic, and non-Bantu language families in general. Nonetheless, our findings are pertinent to a diverse group of populations across Africa. We have shown through our analysis that the combination of high coverage WGS and star allele calling bioinformatics pipelines can facilitate pharmacogenomics analyses in these underrepresented ethnolinguistic groups. Secondly, we used mostly short-read WGS data for our analysis. Therefore, computationally-inferred novel alleles for both *CYP2B6* and *CYP2A6* should be interpreted with caution given the difficulties associated with genotyping these genes (Desta et al., 2021; Wassenaar et al., 2016). In the same vein, we were unable to computationally resolve 26 *CYP2B6* diplotypes (Table 6.4) and four *CYP2A6* diplotypes (Table 6.7) in SSA individuals either due to presence of novel core variants that could not be phased or due to potential uncharacterised structural variations that were novel to all the algorithms used in this study. We have reported the background star alleles for these participants and indicated the potential novel allele-defining variations which require further experimental validation. Thirdly, predicted CYP2A6 phenotypes (relating to nicotine metabolism) could not be assigned to SSA participants as the recently developed genetic risk score (El-Boraie et al., 2021) have only been validated in African Americans/Afro-Caribbeans but not resident African populations. For CYP2B6, the predicted phenotype is only applicable to efavirenz metabolism as per CPIC guidelines (Desta et al., 2019) and about 7% of individuals had indeterminate phenotypes due to habouring alleles with unknown function and/or ambiguous diplotypes. Fourthly, in our laboratory validation, *CYP2B6* XL-PCR products for 27 participants failed to barcode during the HiFi sequencing library preparation due to technical challenges that arose from pooling the *CYP2B6* amplicons with those from *CYP2A6* — for which 28 samples failed barcoding. We mitigated this by optimising barcoding for select individual amplicons from samples that had

predicted novel alleles or suballeles.

In conclusion, this study presents an extensive characterisation of the *CYP2B6* and *CYP2A6* pharmacogenomic variation in diverse SSA populations based on analysis of high depth genomes from multiple cohorts. The differences in *CYP2B6* and *CYP2A6* star allele frequencies across SSA and compared to other global populations, and the high number of potential novel alleles in SSA emphasise the need for pharmacogenomic studies across underrepresented populations to bring precision medicine closer to reality in Africa. Future work entailing in-depth allele characterisation for the *CYP2B6* and *CYP2A6* novel alleles not validated in this study and also functional analyses to determine their clinical function would be critical in supporting clinical pharmacogenomics implementation strategies across Africa.

6.6 Author contributions

D.T., Z.L., and S.H. designed the research; S.H. obtained funding, administrative, technical or material support; D.T. analysed the data; D.T. drafted the manuscript; Z.L. and S.H. critically revised the manuscript for important intellectual content; Z.L. and S.H. supervised the research.

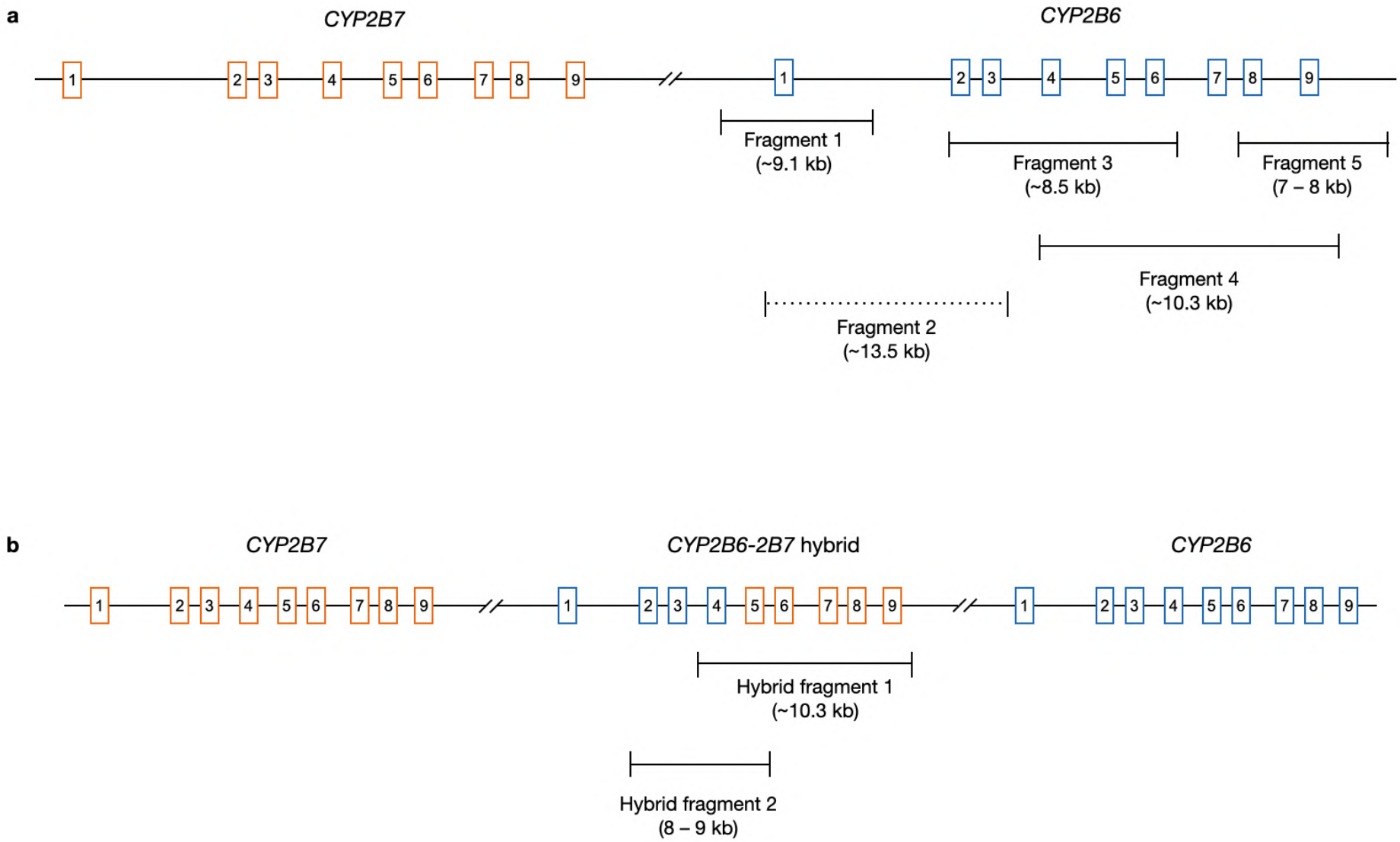


Figure S6.1: *CYP2B6* XL-PCR fragments generated for SMRT sequencing. (**a**) Shows overlapping *CYP2B6* fragments 1–5. Fragment 2 did not amplify in any of our PCR runs. (**b**) Depicts fragments for assaying the presence of a *CYP2B6-2B7* hybrid allele.

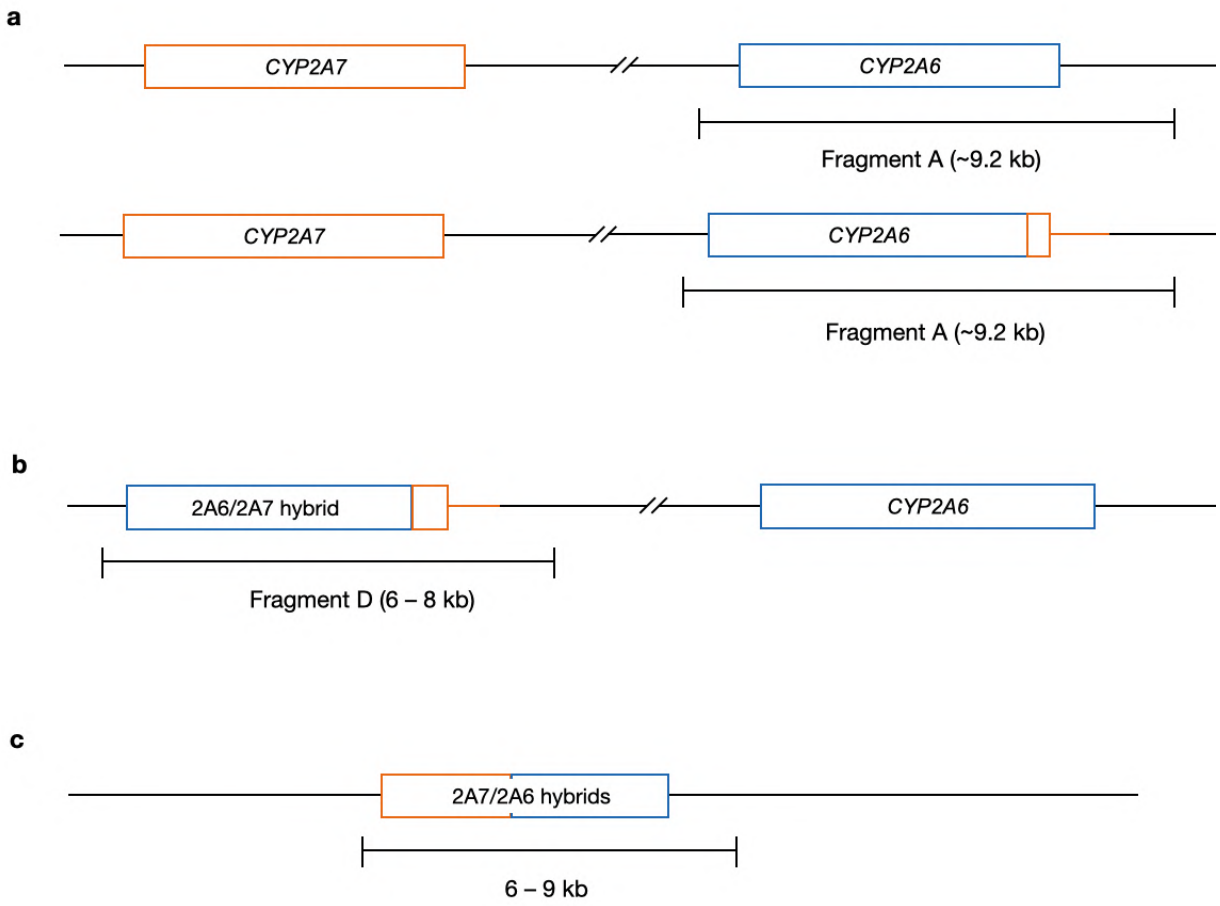


Figure S6.2: *CYP2A6* XL-PCR fragments generated for SMRT sequencing. (**a**) Shows *CYP2A6* fragment A configurations — some comprise the common 3'-UTR gene conversion. (**b**) Depicts the duplication fragment (*CYP2A6-2A7* hybrid) which is a reciprocal of the *CYP2A6*4* deletion. (**c**) Depicts a *CYP2A7-2A6* hybrid fragment (found in *CYP2A6*12* and other alleles).

Table S6.1: Sample information for participants with publicly available high coverage WGS data containing potential novel *CYP2B6* star alleles. The Coriell DNA samples from these participants could potentially be added to GeT-RM reference materials for *CYP2B6* genotyping in pharmacogenomics testing laboratories.

Coriell / SGDP IDs	Predicted consensus diplotype	Country/ population	Primary project
HG02923	*1/[*1 + rs541486480∼5164A>C(K53Q)]	ESN	1000G
HG03163	*6/[*1 + rs541486480∼5164A>C(K53Q)]	ESN	1000G
HG02642	*1/[*1 + rs373926269∼17867T>C(splice donor)]	GWD	1000G
HG03246	*6/[*1 + rs150072531∼26413A>G(H397R)]	GWD	1000G
NA19321	*1/[*6 + rs142421637∼17856C>T(R109W)]	LWK	1000G
NA19334	*6/[*6 + rs142421637∼17856C>T(R109W)]	LWK	1000G
LP6005442-DNA_E11	*6/[*6 + rs142421637∼17856C>T(R109W)]	Kenya	SGDP
NA18916	*1/[*6 + rs58871670∼20669G>A(V183I)]	YRI	1000G
HG03485	*6/[*6 + rs58871670∼20669G>A(V183I)]	ESN	1000G
HG02635	*1/[*6 + rs58871670∼20669G>A(V183I)]	GWD	1000G
HG02763	*17/[*22 + rs141666881∼18137G>A(R158Q)]	GWD	1000G
HG03069	*22/[*22 + rs141666881∼18137G>A(R158Q)]	MSL	1000G
LP6005442-DNA_G10	*1/[*22 + rs141666881∼18137G>A(R158Q)]	Gambia	SGDP
HG02896	*1/[*22 + rs141666881∼18137G>A(R158Q)]	ESN	1000G
HG02464	*1/[*22 + rs141666881∼18137G>A(R158Q)]	GWD	1000G
HG02611	*1/[*22 + rs141666881∼18137G>A(R158Q)]	GWD	1000G
HG03460	*22/[*22 + rs34698757∼23790C>G(T306S)]	MSL	1000G
NA19026	[*1/*6] + rs183427203∼17823C>T(R98W)	LWK	1000G
NA19372	[*6/*17] + rs45459594∼20668C>G (I182M) + rs36079186∼20715T>C (M198T)	LWK	1000G
NA19451	[*1; *22] + possible CYP2B6/2B7 hybrid	LWK	1000G
NA19452	[*1; *22] + possible CYP2B6/2B7 hybrid	LWK	1000G
HG03091	[*1; *22] + possible CYP2B6/2B7 hybrid	MSL	1000G
HG03432	[*1; *22] + possible CYP2B6/2B7 hybrid	MSL	1000G
HG03436	[*1; *22] + possible CYP2B6/2B7 hybrid	MSL	1000G
NA19190	[*1; *22] + possible CYP2B6/2B7 hybrid	YRI	1000G
NA19472	[*1; *22] + possible CYP2B6/2B7 hybrid	YRI	1000G
SS6004471	[*1; *22] + possible CYP2B6/2B7 hybrid	Congo	SGDP
HG02953	[*1; *22] + possible CYP2B6/2B7 hybrid	ESN	1000G
HG02799	[*1; *22] + possible CYP2B6/2B7 hybrid	GWD	1000G
HG02813	[*1; *22] + possible CYP2B6/2B7 hybrid	GWD	1000G
LP6005443-DNA_F02	[*2/*9] + rs28399499∼26018T>C(I328T)	Botswana/ Namibia	SGDP

cn, copy number; fs, frameshift; 1000G, 1000 Human Genomes Project; SGDP, Simons Genome Diversity Project; ESN, Esan in Nigeria; LWK, Luhya in Webuye Kenya; GWD, Gambian from Western Divisions (Mandinka); MSL, Mende in Sierra Leone; YRI, Yoruba in Ibadan, Nigeria; GeT-RM, Genetic Testing Reference Materials Coordination Program.

Table S6.2: Sample information for participants with publicly available high coverage WGS data containing potential novel *CYP2B6* star alleles.

Coriell / SGDP IDs	Predicted consensus diplotype	Country/ population	Primary project
HG02582	*1/[*1 + rs554865113∼6727G>C]	ESN	1000G
HG02811	[*1 + rs114558780∼10126C>T]/[*1 + rs2545783∼6798G>T]; cn=3	ESN	1000G
HG03297	*1/[*1 + rs557976670∼8433C>T]	LWK	1000G
NA19393	*1/[*1 + rs138978736∼8457C>A]	YRI	1000G
HG02981	*17/[*1 + rs138978736∼8457C>A]	GWD	1000G
NA18517	*1/[*1 + rs58720852∼8515C>A]	MSL	1000G
NA19025	*1/[*1 + rs111869995∼8567G>T]	MSL	1000G
HG02465	*1/[*1 + rs528089983∼9445C>T]	ESN	1000G
HG02573	*1/[*1 + rs528089983∼9445C>T]	ESN	1000G
HG03461	*1/[*1 + rs58571639∼9450C>T]	LWK	1000G
NA19440	*1/[*1 + rs145036049∼10108G>A]	YRI	1000G
NA19451	*1/[*1 + rs114558780∼10126C>T]	YRI	1000G
NA19452	*1/[*1 + rs114558780∼10126C>T]	YRI	1000G
SS6004471	*1/[*1 + rs114558780∼10126C>T]	Congo	SGDP
HG03054	*9/[*1 + rs114558780∼10126C>T]	GWD	1000G
HG03081	*1/[*1 + rs114558780∼10126C>T]	GWD	1000G
HG03479	*1/[*1 + rs114558780∼10126C>T]	LWK	1000G
NA18912	*1/[*1 + rs114558780∼10126C>T]	MSL	1000G
HG02611	*1/[*1 + rs28399463∼10766A>G]	ESN	1000G
HG02851	*1/[*1 + rs28399463∼10766A>G]	ESN	1000G
HG02861	*1/[*1 + rs28399463∼10766A>G]	ESN	1000G
HG02881	*1/[*1 + rs28399463∼10766A>G]	ESN	1000G
NA19435	*1/[*1 + rs8192730∼10771G>C]	YRI	1000G
SS6004473	[*1 + rs113558392∼9394T>G]/ [rs533216061∼8508C>A + rs771986786∼7227C>G]	Namibia	SGDP
LP6005441-DNA_A11	[(*1 + rs113558392∼9394T>G) x2]/[(*1 + rs113558392∼9394T>G) x2]	Namibia	SGDP
LP6005443-DNA_G08	*31/[(*1 + rs113558392∼9394T>G) x2]	Namibia	SGDP
LP6005592-DNA_C05	*9/[(*1 + rs113558392∼9394T>G) x2]	South Africa	SGDP
LP6005677-DNA_D03	*1/[(*1 + rs113558392∼9394T>G) x2]	South Africa	SGDP
NA19467	*1/[*1 + rs61605570∼9990A>T]	YRI	1000G
NA18522	[*9/*27] + rs143731390∼11479A>T	MSL	1000G
NA19035	*17/[*9 + rs145308399∼5576G>A]	MSL	1000G
NA19308	*1/[*9 + rs145308399∼5576G>A]	YRI	1000G
HG03129	*4/[*17 + rs554920226∼10778C>A]	GWD	1000G
NA19375	*1/[*18 + rs4997557∼9400C>G + rs2644906∼9393G>A]	YRI	1000G
SS6004480	*1/[*18 + rs4997557∼9400C>G + rs2644906∼9393G>A]	South Sudan	SGDP
HG03079	*17/[*18 + rs4997557∼9400C>G + rs2644906∼9393G>A]	GWD	1000G
HG02757	[*9/*35] + rs137904044∼9983G>T	ESN	1000G
HG03470	[*17/*35] + rs143067113∼5546G>A	LWK	1000G
NA18522	[*9/*27] + rs143731390∼11479A>T	MSL	1000G
LP6005443-DNA_B09	[*9/*17] + unresolved CYP2A6 duplication	South Sudan	SGDP

cn, copy number; fs, frameshift; 1000G, 1000 Human Genomes Project; SGDP, Simons Genome Diversity Project; ESN, Esan in Nigeria; LWK, Luhya in Webuye Kenya; GWD, Gambian from Western Divisions (Mandinka); MSL, Mende in Sierra Leone; YRI, Yoruba in Ibadan, Nigeria; GeT-RM, Genetic Testing Reference Materials Coordination Program.

Chapter 7

Discussion and Conclusion

This PhD study aimed to characterise the genetic variation in three key pharmacogenes i.e. *CYP2D6*, *CYP2B6* and *CYP2A6* in African populations. The strengths and limitations of existing pharmacogene star allele calling tools are also examined herein in addition to the presentation of a novel diplotype calling algorithm. The study objectives were achieved through four empirical studies presented as individual chapters within this thesis. Table 7.1 provides a summary of the key findings from each of these chapters. This final chapter discusses the the overarching research themes from this PhD work in the African context, highlights the strengths and limitations of the study, and also outlines potential future studies.

Table 7.1: Consolidated findings from the four empirical studies included in this thesis.

Objective 1: To compare the performance of existing pharmacogene star allele calling algorithms using HTS data.

Chapter: 3

Main findings:

- Diplotype concordance ranged from 71% (Astrolabe) to 88% (Stargazer) among existing tools when used on real-world data.
- Astrolabe had lower recall for *CYP2D6* CNVs and hybrid rearrangements compared to Aldy and Stargazer.
- Stargazer had low recall for rare alleles, especially in heterozygous combinations, due to reliance on statistical phasing.
- Aldy had the highest accuracy for calling *CYP2D6* structural variations; however, it called some false positive **61*, **63* and **13* alleles even in samples without structural variants.
- All three tools were prone to ambiguous calls and incorrect calls if alleles/suballeles were not comprehensively defined in their databases.
- The ensemble genotyping approach had comparably high *CYP2D6* haplotype and diplotype concordance to the truth set, but it was prone to ambiguous calls especially for samples with structural variations — where the three tools often reported different calls.
- The ensemble approach (i.e. using at least 3 algorithms) is recommended for future application of *in silico CYP2D6* genotyping in clinical and research settings.

Table 7.1 – continued

Objective 2: To develop a novel graph- and Nextflow-based algorithm for calling star alleles in highly polymorphic and structurally variant pharmacogenes.

Chapter: 4

Main findings:

– We developed StellarPGx which assigns star alleles using a hybrid approach involving genome graph-based variant calling, combinatorial allele discrimination, and computing read coverage information.
– StellarPGx's *CYP2D6* allele calling performance was comparable that for Cyrius, and superior to Aldy and Stargazer based on the analysis of WGS data corresponding to the GeT-RM reference materials.
– StellarPGx had a shorter runtime compared to Cyrius during single and multi-sample analyses.
– StellarPGx can call diplotypes in 12 CYP genes while Cyrius is currently only available for *CYP2D6*.
– StellarPGx's diplotype concordance to the GeT-RM reference materials was lower for *CYP1A1*, *CYP2E1*, and *CYP2B6* compared to other supported CYP genes.
– More comprehensive allele definitions and reference materials are needed for all core CYP genes to enhance the allele calling ability of the available computational tools.

Objectives 3 and 4:

– To identify common, rare, and potential novel alleles in *CYP2D6*, *CYP2B6*, and *CYP2A6* in African populations using WGS data, and predict the corresponding phenotype distribution.
– To validate novel African-specific alleles and resolve ambiguous genotypes in *CYP2D6*, *CYP2B6*, and *CYP2A6* by targeted long-read resequencing.

Chapter: 5 (*CYP2D6*)

Main findings:

– Our *CYP2D6* star allele analysis revealed non-uniform allele distributions across the different SSA populations in this study.
– We observed unique distributions of the CYP2D6 metaboliser phenotypes across different SSA populations and also in comparison to African American/Afro-Caribbean, European, admixed American, East Asian and South Asian populations.
– We identified 27 high-confidence potential novel *CYP2D6* alleles in SSA populations, 19 of which were fully phased (computationally).
– HiFi sequencing enabled the validation of two predicted novel major alleles (Figure 5.8 on page 140 and Figure 5.9 on page 141) in the first two sample batches sent for targeted resequencing and revealed novel suballeles in multiple participants.

Table 7.1 – continued

Chapter: 6 (*CYP2B6* and *CYP2A6*)

Main findings:

- We found largely distinct *CYP2B6* and *CYP2A6* allele distributions across the different SSA populations in this study.
- The *CYP2B6*6* allele defined by rs3745274 (Q172H) and rs2279343 (K262R) was the most frequent decreased function *CYP2B6* allele in SSA participants, with an allele frequency was non-uniform ranging from 21.7% in the Batswana to 47% among individuals from Burkina Faso.
- We identified 19 predicted novel *CYP2B6* haplotypes in SSA populations. Fourteen of these alleles were fully phased computationally while five were in ambiguous diplotypes.
- We identified 31 predicted novel *CYP2A6* haplotypes in SSA populations. Twenty eight of these alleles were fully phased computationally while the other three occurred in unresolved diplotypes.
- For *CYP2B6* and *CYP2A6*, HiFi sequencing enabled the validation of two predicted novel major alleles each in the first two batches of samples sent for targeted resequencing.

7.1 Emerging Research Themes

The specific results in this PhD study have been discussed in detail in the four component empirical studies. The pharmacogenomics aspects investigated in this thesis can be broadly divided into three research themes i.e. (i) the development, benchmarking and application of star allele calling algorithms (ii) the characterisation of *CYP2D6*, *CYP2B6* and *CYP2A6* allelic variation in African populations and (iii) the validation of new pharmacogenetic alleles using long-read sequencing.

7.1.1 Development, benchmarking and application of star allele calling algorithms

Chapter 3 examined the strengths and limitations of existing pharmacogene star allele callers, and provided a much-needed benchmark as well as recommendations regarding computational-based *CYP2D6* genotyping, based on both simulated and real-world HTS data. Chapter 4 presented StellarPGx, a novel pipeline for calling star alleles in cytochrome P450 genes which employs a hybrid approach involving pangenome-based variant detection, combinatorial diplotype assignment, and evaluation of read depth information from WGS data.

Accurately determining the phase of allele-defining SNVs in key pharmacogenes is challenging, in particular when analysing short-read HTS data. Despite their often high accuracy, short reads do not cover the entire gene locus, meaning that haplotypes cannot be determined directly from short-read HTS data. Furthermore, for pharmacogenes with neighbouring paralogous pseudogenes e.g. *CYP2D6* which lies alongside the *CYP2D7* and *CYP2D8* pseudogenes, short reads can misalign between highly similar regions. This can lead to erroneous SNV calling, and also complicates the task of structural variant detection especially for individuals carrying hybrid alleles and/or tandem rearrangements (Yang et al., 2017; Turner et al., 2021). However, the ability to tease out pharmacogene diplotypes from short-read HTS data is critical as short-read NGS is ubiquitous, relatively inexpensive and therefore more readily accessible compared to other NGS technologies.

The bioinformatics star allele calling algorithms benchmarked in Chapter 3 and Chapter 4 i.e. Astrolabe, Aldy, Stargazer, Cyrius and StellarPGx (developed in this study) provide an approach for clinicians and researchers to analyse short-read HTS data for purposes of obtaining baseline frequencies of alleles in key ADME genes, and automating phenotype predictions. However, as the algorithms above vary in star allele calling accuracy (according to our benchmarking study), the key recommendation arising from this thesis is to use at least three of the available tools when genotyping some of the more challenging pharmacogenes e.g. *CYP2D6*, *CYP2B6*, and *CYP2A6*. The ensemble genotyping approach is particularly valuable when analysing genomes from understudied populations as these are more likely to contain new allelic variants. In particular, StellarPGx which was developed during this PhD study provides an enhanced ability to identify individuals with potential novel alleles for inclusion in follow-up allele validation studies using unequivocal techniques (Gaedigk, Turner, et al., 2019). Furthermore, the implementation of StellarPGx in Nextflow (Di Tommaso et al., 2017) facilitates pipeline best practices including reproducibility and portability. For example, StellarPGx can easily be adapted for pharmacogenetic analysis of WGS data in cloud-computing environments such as the Amazon Web Services (AWS) resources (`https://aws.amazon.com`).

With regards to tool development, the specific results detailed in Chapter 3 and Chapter 4 provide important insights to researchers that are interested in enhancing the existing allele callers and/or developing new ones as bioinformatics techniques continue to advance. Just as the advent of genome graph-based variant detection approaches led to the idea of developing StellarPGx, more algorithms based on key bioinformatics advancements might help in improving the accuracy of genotyping challenging pharmacogenes such as *CYP2D6*.

One of the next steps in precision medicine in Africa and worldwide is to apply the aforementioned algorithms more regularly in clinical settings so as to promote timely allele detection and phenotype predictions which are needed for adjusting therapy. However, more work is needed in improving

the accuracy of all existing algorithms to a level where high-confidence allele calls can be obtained in any ADME gene as a basis of clinical decisions. In particular, it is important to establish the minimum quality and coverage of HTS data required as input to the in silico tools. In addition, it is imperative for continuous updates to be made to the databases of existing tools as more alleles are submitted to PharmVar. As observed in Chapter 3, the re-definition of alleles and suballeles in the databases of Aldy and Astrolabe markedly improved their *CYP2D6* allele calling accuracy.

7.1.2 Characterisation of *CYP2D6*, *CYP2B6* and *CYP2A6* allelic variation in African populations

Given that there are significant genotype-related differences in the response to medications commonly administered across Africa (Adedeji et al., 2017; Dodgen et al., 2016; Maimbo et al., 2012; Pernaute-Lau et al., 2021; Petros et al., 2017; van der Merwe et al., 2012), and that both communicable and non-communicable diseases are on the rise, clinical pharmacogenomics implementation is critical across the continent. However, Africa has been under-represented in pharmacogenomics research to date (Desta et al., 2021; Gaedigk et al., 2017; Rajman et al., 2017), hence information about the distribution of pharmacogenetic variants and/or haplotypes across most African ethnolinguistic groups is scarce. The increasing availability of HTS datasets from diverse African populations and the availability of bioinformatics tools to use for mining these datasets provided the opportunity to address this gap in knowledge for important pharmacogenes i.e. *CYP2D6*, *CYP2B6* and *CYP2A6*. CYP2D6 is involved in the bioactivation or elimination of over 25% of commonly prescribed medications, while CYP2B6 and CYP2A6 play key roles in the metabolism of antiretroviral medications and nicotine.

In line with previous pharmacogenomics studies involving African participants (Matimba et al., 2009; Rajman et al., 2017; Wright et al., 2010), this study showed that there are non-uniform distributions of known alleles in *CYP2D6*, *CYP2B6* and *CYP2A6* across populations in SSA based on the analysis of 962 high depth African genomes. The bulk of these data was primarily generated by the H3African Consortium and the 1000 Genomes Project Consortium, and many of the genomes were from previously undersampled populations across central, east, west, and southern Africa. The observed allele frequency differences underline the importance of investigating pharmacogenetic variation in the African context and in diverse ethnolinguistic groups to support the development of effective precision medicine strategies across Africa. This approach is also supported by the numerous potential novel African-specific *CYP2D6*, *CYP2B6* and *CYP2A6* alleles predicted in this study. During the short-read WGS data analysis, 27 potential novel *CYP2D6* alleles were observed across SSA, while 19 and 31 potential novel alleles were observed for *CYP2B6* and *CYP2A6*, respectively. Given the aforementioned comp-

lexity associated with genotyping *CYP2D6*, *CYP2B6* and *CYP2A6*, these predicted novel alleles should be interpreted with caution, especially those that have not undergone further experimental validation. All these potential novel alleles were rare which, in addition to the observed allele frequency differences, supports the fact that one African population, let alone African Americans/ Afro-Caribbeans, cannot be used as a proxy for the whole African continent when developing clinical pharmacogenomics implementation strategies.

In addition to the star allele analysis, this study investigated the distribution of predicted CYP2D6 and CYP2B6 phenotypes across SSA, based on the inferred diplotypes and the activity score system (Caudle et al., 2020; Desta et al., 2019). The phenotype distributions were also non-uniform across the SSA populations in this study which typifies the diversity that needs to be taken into account for effective precision medicine implementation across Africa. Metaboliser phenotypes were not predicted for CYP2A6 as the genetic risk score that has been developed by El-Boraie et al. (2021) for African-ancestry individuals has only been validated in African American and Yoruban participants.

7.1.3 Validation of new pharmacogenetic alleles using long-read sequencing

Experimental validation of novel *CYP2D6* star alleles often involves using a battery of techniques such as XL-PCR, allele-specific PCR, pharmacogenetic arrays and panels, and Sanger sequencing (Gaedigk, Turner, et al., 2019; Pratt et al., 2016). This is because of the aforementioned complexities in genotyping *CYP2D6* caused mainly by the neighbouring highly similar pseudogenes. However, Sanger sequencing is expensive, laborious and limited in the length of reads it can produce, while allele-specific PCR might require multiple rounds of optimisation. Targeted long-read sequencing has recently been established as a reliable tool for *CYP2D6* star allele characterisation by researchers using either Oxford Nanopore (Liau et al., 2019; Ammar et al., 2015) or Pacific Biosiences (Qiao et al., 2016; Buermans et al., 2017) platforms. Long reads can span the entire length of smaller pharmacogenes e.g. *CYP2D6* and *CYP2A6*, thus facilitating simultaneous variant phasing and ascertainment of duplicated gene copies. For larger genes such as *CYP2B6* and *CYP2C19*, long reads generated from overlapping fragments can also offer similar advantages.

This study applied targeted long-read SMRT sequencing to characterise some of the novel African-specific alleles predicted in *CYP2D6*, *CYP2B6* and *CYP2A6*. The amplicons of interest were generated via XL-PCR and then SMRT sequencing was done on the Sequel IIe platform at a Pacific Biosciences certified service center (Inqaba Biotec, South Africa). The Sequel IIe platforms generates HiFi reads which are both long (up to 15kb) and accurate (>99.9%) making them suitable for characterising alleles in these complex pharmacogenes. In particular, this study is the first to use targeted SMRT sequencing to characterise novel alleles in the *CYP2A6* gene

(see subsubsection 6.4.5), and also to employ multiplex HiFi sequencing for *CYP2D6*, *CYP2B6* and *CYP2A6* amplicons.

7.2 Strengths and limitations

This PhD study had its strengths and also some limitations. The strengths include (i) the diversely sampled study population representative of over 13 ethnolinguistic groups across SSA (ii) the use of high depth genomes from which both SNVs and structural variants could be called with a high level of confidence (iii) the high accuracy and improved features of the more recently developed star allele calling algorithms i.e. Cyrius and StellarPGx — which was developed in this study (iv) the combination of XL-PCR and HiFi sequencing for the validation of predicted novel star alleles (v) the application of these pharmacogenetic research strategies in the African context.

The limitations of the individual study components have been discussed within each of the chapters presented in manuscript format. The broader limitations discussed in this section include (i) the uneven sample sizes, including virtually no representation of northern African populations (ii) the inability to resolve all diplotypes in the full dataset (iii) the inability to assign predicted phenotypes corresponding to all the assigned individual diplotypes (iv) limitations relating to the laboratory validation of predicted novel alleles (v) limitations relating to the currently available allele definitions.

With reference to the sampling, it is important to acknowledge that there was uneven representation of regions across Africa e.g. there were more samples included from southern Africa compared to those from eastern and far western Africa, and we had WGS data from only one northern African participant. This was in part because this study analysed secondary genomic data and did not involve new sample collection. In terms of ethnolinguistic groups, we had very few samples from Nilo-Saharan and Afroasiatic language speakers, and a limited number of samples from speakers of non-Bantu languages SSA such as the Khoe and San speakers. Given the findings discussed in Chapter 5 and Chapter 6, it is important to include more sparsely sampled populations to date in pharmacogenomics studies across Africa to obtain a more comprehensive pharmacogene variation landscape (da Rocha, Othman, Botha, et al., 2021), rather than attempting to extrapolate findings from other populations.

With regards to the inability to resolve all the diplotypes in the full dataset, this is mainly related to the use of the ensemble genotyping approach where the tools used reported completely different star alleles for a few of the samples. In addition, for individuals with novel alleles occurring together with heterozygous background alleles, it was difficult to infer the phase of the novel core SNVs, in particular for the participants not included in the laboratory validation due to lack of available DNA samples. Furthermore, given that complex structural variations are known

to occur in *CYP2D6*, *CYP2B6* and *CYP2A6*, the occurrence of ambiguous diplotypes exposes the aforementioned limitations of using short-read HTS data for analysing these complex loci. For some of the participants, manual inspection of the read data provided better resolution in terms of obtaining the most-likely diplotype call. However, the unequivocal characterisation of the alleles carried by these participants using validated experimental approaches is perhaps the more robust way of ensuring that later versions of the current algorithms can call them consistently.

The phenotype prediction approach used in this study for CYP2D6 and CYP2B6 is in line with CPIC recommendations (Desta et al., 2019). However, there was a considerable number of individuals to whom an indeterminate metaboliser status was assigned for one or more of the following reasons: (i) the diplotype was ambiguous or indeterminate (ii) one of the alleles in the diplotype had an unknown or uncertain function (iii) the individual had a predicted novel allele. Furthermore, no phenotype prediction was done for CYP2A6 as there is no equivalent activity score and/or weighted genetic risk score validated for resident African populations except for the Yoruban participants in the 1000 Genomes Project (El-Boraie et al., 2021). In addition, as reviewed in Chapter 2, other factors are known to affect the phenotypes of CYP2D6, CYP2B6 and CYP2A6 including (i) phenocoversion i.e. where an individual presents with a metaboliser phenotype contrary to that predicted from their genotype due to inhibition or induction of the metabolising enzyme by other substrate(s) e.g. other medications and herbal remedies (ii) long-range enhancer polymorphisms (iii) substrate specificity of ADME proteins — therefore, it is important to account for these factors when applying precision medicine strategies based on the CYP2D6 and CYP2B6 phenotypes predicted in this study.

With regards to the laboratory validation of predicted novel alleles, two novel major African-specific alleles apiece were characterised in *CYP2D6*, *CYP2B6* and *CYP2A6*, in addition to multiple novel suballeles. However, there were shortcomings in this validation work: (i) insufficient DNA for exhaustive validation of novel alleles or suballeles for some participants e.g. *CYP2D6-2D7* hybrid assays were omitted for most of the samples while *CYP2D6* FragA assays were prioritised depending on the *in silico* predictions (see Chapter 5) (ii) lack of access to DNA samples for some of the individuals predicted to have novel alleles (iii) uneven SMRT sequencing quality of multiplexed *CYP2D6*, *CYP2B6* and *CYP2A6* amplicons. For some of the samples, application of other techniques such as digital droplet PCR and gene deletion assays would have been useful to conclusively determine the copy number states for *CYP2D6*, *CYP2B6* and *CYP2A6*. Overall, the main focus of the amplification strategies was to generate fragments for targeted long-read SMRT sequencing. The *in silico* star allele calling that preceded the experimental validation work was useful in prioritising samples for targeted SMRT sequencing thereby mitigating some of the aforementioned challenges. Another related challenge was the depletion of sample DNA while optimising the generation of XL-PCR fragments from the more

complex and rare allelic arrangements. Targeted SMRT sequencing for *CYP2B6* was particularly challenging as *CYP2B6* had more larger sized amplicons 10-13.5 kb that were either difficult to amplify or sequenced insufficiently alongside the smaller *CYP2D6* and *CYP2A6* fragments.

With regards to the allele definitions, the different allele calling tools used in this study relied largely on information from versions 4.0.4 - 5.0 of the PharmVar database. It is possible that some diplotypes might have been assigned differently by some of the tools if the novel alleles or suballeles predicted in this study — and also those yet to be identified and submitted to PharmVar by other studies — had been added to the databases of these tools. In particular, the *CYP2A6* allele definitions were not updated by PharmVar during the course of this analysis. PharmVar's *CYP2A6* panel, which I have recently joined as a member, is currently undertaking the curation of all *CYP2A6* allele definitions — this means that some existing haplotype definitions could be retired while others might be assigned new star numbers in the near-future. The *CYP2A6* allele calling results from this study are therefore to be interpreted according to the available nomenclature before the expected roll out of the new *CYP2A6* allele definitions by PharmVar.

7.3 Future work

This study has generated a number of interesting findings which have opened up future research directions. Subsequent studies that could build on this work are outlined below:

Calling star alleles in other key pharmacogenes

The increasing availability of WGS data from understudied populations and the bioinformatics tools benchmarked in Chapter 3 and Chapter 4 provide the opportunity to investigate the star allele diversity in key pharmacogenes that code for drug metabolising enzymes other than CYP2D6, CYP2B6 and CYP2A6, in the African context. These include genes such as *CYP2C19*, *CYP2C9*, *CYP2C8*, and *TPMT*. In addition, it would be interesting to investigate the genetic variation in genes that encode drug transporter proteins such as SLCO1B1. Given the relatively high number of potential novel African-specific alleles identified in *CYP2D6*, *CYP2B6* and *CYP2A6*, the analysis of alleles in other key pharmacogenes promises to reveal more uncharacterised pharmacogenomically-relevant haplotypes.

Star allele analysis for understudied African populations not included in this study

As mentioned in the study limitations, our analysis of secondary data lacked adequate representation of northern African populations and non-Bantu ethnolinguistic groups across SSA. Future studies involving these populations are therefore warranted in order to obtain an accurate landscape of pharmacogene variation across Africa and also identify more rare and/or novel alleles. This information will be vital for the development of comprehensive genotyping panels with better utility across Africa.

Experimental validation of the full set of predicted novel alleles

As presented on Chapter 5 and Chapter 6, this study predicted over 50 novel alleles in *CYP2D6*, *CYP2B6* and *CYP2A6* combined. Of these, only six novel major alleles were validated in this study. However, given the complexity relating to genotyping these three genes (see Chapters 2-6), it is imperative to perform experimental validation for the full set of predicted novel alleles before they can be added to clinical pharmacogenomics testing platforms. More of these predicted novel alleles are expected to be characterised using HiFi reads when targeted SMRT sequencing of additional sample batches is complete. Furthermore, validation of predicted novel *CYP2D6* alleles occurring in participants with corresponding DNA samples available from the Corielle Institute (Camden, USA) is already underway as part of an ongoing collaboration with PharmVar.

Enhancements to StellarPGx

StellarPGx (described in Chapter 4) is one of the major contributions of this thesis to the pharmacogenomics and precision medicine communities in Africa and around the world. By facilitating accurate star allele detection from short-read WGS data, identification of individuals with potential novel haplotypes and also automating phenotype prediction for genes with a standardised activity system, StellarPGx can potentially be an important bioinformatics tool for routine use in both research and clinical pharmacogenomics implementation settings. However, as StellarPGx currently uses only short-read WGS as input, future enhancements are necessary regarding diversifying the input data types to WES, long-read WGS, ADME panel data and possibly array genotype data (for less complex pharmacogenes). Furthermore, all the alleles under validation in ongoing and future collaborations, including those submitted to PharmVar by other research groups need to be added into the StellarPGx database. Additional enhancements to StellarPGx entail making the tool as easy to use as possible e.g. embedding StellarPGx algorithms into web-based platforms such as Galaxy and documenting how to run it in cloud computing environments.

In conclusion, this PhD study has highlighted the distribution of common, rare and/or novel *CYP2D6*, *CYP2B6*, and *CYP2A6* star alleles in multiple resident African populations, thus contributing to bridging the knowledge gaps hindering the development and application of precision medicine strategies across Africa. In addition, this study has contributed a highly relevant and useful allele calling bioinformatics tool (StellarPGx) to the pharmacogenomics community, with the possibility for application in clinical settings especially after the aforementioned enhancements.

References

Adedeji, W. A., Igbinoba, S. I., Fakeye, T. O., Oladosu, I. A., Fehintola, F. A., Ma, Q., and Morse, G. D. (2017). Evaluation of CYP2D6 phenotype in the Yoruba Nigerian population. *Expert Review of Clinical Pharmacology*, **10**(10), 1145–1152.

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in Human Genetics*, **Chapter 7**, Unit7.20.

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., et al. (2021). A complete reference genome improves analysis of human genetic variation. *bioRxiv*, 2021.07.12.452063.

Ahmed, J. H., Makonnen, E., Fotoohi, A., Aseffa, A., Howe, R., and Aklillu, E. (2019). *CYP2D6* genotype predicts plasma concentrations of tamoxifen metabolites in Ethiopian breast cancer patients. *Cancers*, **11**(9), 1353.

Aklillu, E., Persson, I., Bertilsson, L., Johansson, I., Rodrigues, F., and Ingelman-Sundberg, M. (1996). Frequent distribution of ultrarapid metabolizers of debrisoquine in an Ethiopian population carrying duplicated and multiduplicated functional *CYP2D6* alleles. *The Journal of Pharmacology and Experimental Therapeutics*, **278**(1), 441–446.

Al Koudsi, N., Hoffmann, E. B., Assadzadeh, A., and Tyndale, R. F. (2010). Hepatic *CYP2A6* levels and nicotine metabolism: impact of genetic, physiological, environmental, and epigenetic factors. *European Journal of Clinical Pharmacology*, **66**(3), 239–251.

Alessandrini, M., Asfaha, S., Dodgen, T. M., Warnich, L., and Pepper, M. S. (2013). Cytochrome P450 pharmacogenetics in African populations. *Drug Metabolism Reviews*, **45**(2), 253–275.

Alessandrini, M., and Pepper, M. S. (2014). Priority pharmacogenetics for the african continent: focus on CYP450. *Pharmacogenomics*, **15**(3), 385–400.

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

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, **21**(1), 30.

Ambler, J. M., Mulaudzi, S., and Mulder, N. (2019). Gengraph: a python module for the simple generation and manipulation of genome graphs. *BMC Bioinformatics*, 519.

Ammar, R., Paton, T. A., Torti, D., Shlien, A., and Bader, G. D. (2015). Long read nanopore sequencing for detection of *hla* and *cyp2d6* variants and haplotypes. *F1000Research*, **4**, 17.

Ampadu, H. H., Hoekman, J., Arhinful, D., Amoama-Dapaah, M., Leufkens, H., and Dodoo, A. (2018). Organizational capacities of national pharmacovigilance centres in Africa: assessment of resource elements associated with successful and unsuccessful pharmacovigilance experiences. *Globalization and Health*, **14**, 109.

Ampadu, H. H., Hoekman, J., de Bruin, M. L., Pal, S. N., Olsson, S., Sartori, D., et al. (2016). Adverse drug reaction reporting in Africa and a comparison of individual case safety report characteristics between Africa and the rest of the world: Analyses of spontaneous reports in VigiBase®. *Drug Safety*, **39**(4), 335–345.

Applied Biosystems. (2009). Retrieved June 20, 2021, from `https://tools.thermofisher.com/`

content/sfs/manuals/cms_041003.pdf.

Ariyoshi, N., Takahashi, Y., Miyamoto, M., Umetsu, Y., Daigo, S., Tateishi, T., et al. (2000). Structural characterization of a new variant of the *CYP2A6* gene (*CYP2A6*1B*) apparently diagnosed as heterozygotes of *CYP2A6*1A* and *CYP2A6*4C*. *Pharmacogenetics*, **10**(8), 687–693.

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Awandu, S. S., Raman, J., Makhanthisa, T. I., Kruger, P., Frean, J., Bousema, T., ... Birkholtz, L. M. (2018). Understanding human genetic factors influencing primaquine safety and efficacy to guide primaquine roll-out in a pre-elimination setting in southern Africa. *British Journal of Clinical Pharmacology*, **17**(1), 120.

Bahar, M. A., Kamp, J., Borgsteede, S. D., Hak, E., and Wilffert, B. (2018). The impact of CYP2D6 mediated drug–drug interaction: a systematic review on a combination of metoprolol and paroxetine/fluoxetine. *British Journal of Clinical Pharmacology*, **84**(12), 2704–2715.

Bathum, L., Skjelbo, E., Mutabingwa, T. K., Madsen, H., Hørder, M., and Brøsen, K. (1999). Phenotypes and genotypes for CYP2D6 and CYP2C19 in a black Tanzanian population. *British journal of Clinical Pharmacology*, **48**(3), 395–401.

Bentley, A. R., Callier, S. L., and Rotimi, C. N. (2020). Evaluating the promise of inclusion of African ancestry populations in genomics. *npj Genomic Medicine*, **5**(1), 1–9.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218), 53–59.

Berhe, D. F., Juhlin, K., Star, K., Beyene, K. G. M., Dheda, M., Haaijer-Ruskamp, F. M., et al. (2015). Adverse drug reaction reports for cardiometabolic drugs from sub-Saharan Africa: a study in VigiBase. *Tropical Medicine and International Health*, **20**(6), 797–806.

Black, J. L., Walker, D. L., O'Kane, D. J., and Harmandayan, M. (2012). Frequency of undetected *CYP2D6* hybrid genes in clinical samples: impact on phenotype prediction. *Drug Metabolism and Disposition*, **40**(1), 111–119.

Borba, M. A., Melo-Neto, R. P., Leitão, G. M., Castelletti, C. H., Lima-Filho, J. L., and Martins, D. B. (2016). Evaluating the impact of missenses mutations in *CYP2D6*7* and *CYP2D6*14A*: does it compromise tamoxifen metabolism? *Pharmacogenomics*, **17**(6), 561–570.

Breese, M. R., and Liu, Y. (2013). NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics*, **29**(4), 494–496.

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, **81**(5), 1084–1097.

Buermans, H. P. J., Vossen, R. H. A. M., Anvar, S. Y., Allard, W. G., Guchelaar, H.-J., White, S. J., et al. (2017). Flexible and scalable full-length *CYP2D6* long amplicon PacBio sequencing. *Human Mutation*, **38**(3), 310–316.

Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., et al. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nature Communications*, **11**(1), 6272.

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project, cohort including 602 trios. *bioRxiv*, 2021.02.06.430068.

Cai, W.-M., Nikoloff, D. M., Pan, R.-M., de Leon, J., Fanti, P., Fairchild, M., … Wedlund, P. J. (2006). *CYP2D6* genetic variation in healthy adults and psychiatric African-American subjects: implications for clinical practice and genetic testing. *The Pharmacogenomics Journal*, **6**(5), 343–350.

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**, 403–433.

Cao, H., Hastie, A. R., Cao, D., Lam, E. T., Sun, Y., Huang, H., et al. (2014). Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience*, **3**(1).

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, **14**(Suppl 3), S3.

Caudle, K. E., Gammal, R. S., Whirl-Carrillo, M., Hoffman, J. M., Relling, M. V., and Klein, T. E. (2016). Evidence and resources to implement pharmacogenetic knowledge for precision medicine. *American Journal of Health-System Pharmacy*, **73**(23), 1977–1985.

Caudle, K. E., Sangkuhl, K., Whirl-Carrillo, M., Swen, J. J., Haidar, C. E., Klein, T. E., et al. (2020). Standardizing *CYP2D6* genotype to phenotype translation: Consensus recommendations from the clinical pharmacogenetics implementation consortium and dutch pharmacogenetics working group. *Clinical and Translational Science*, **13**(1), 116–124.

Chen, X., Shen, F., Gonzaludo, N., Malhotra, A., Rogert, C., Taft, R. J., … Eberle, M. A. (2021). Cyrius: accurate *cyp2d6* genotyping using whole-genome sequencing data. *The Pharmacogenomics Journal*, **21**(2), 251–261.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLOS ONE*, **7**(10), e46688.

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., et al. (2020). High-depth African genomes inform human migration and health. *Nature*, **586**(7831), 741–748.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., et al. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, **8**(1), 2062.

Chun, S., and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research*, **19**(9), 1553–1561.

Collins, R. L., Brand, H., Karczewski, K. J., et al. (2020). A structural variation reference for medical and population genetics. *Nature*, **581**, 444–451.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, **38**(11), 1251–1260.

Croom, E. L., Stevens, J. C., Hines, R. N., Wallace, A. D., and Hodgson, E. (2009). Human hepatic CYP2B6 developmental expression: the impact of age and genotype. *Biochemical Pharmacology*, **78**(2), 184–190.

Dalton, R., Lee, S.-B., Claw, K. G., Prasad, B., Phillips, B. R., Shen, D. D., et al. (2020). Interrogation of *CYP2D6* structural variant alleles improves the correlation between *CYP2D6* genotype and CYP2D6-mediated metabolic activity. *Clinical and Translational Science*, **13**(1), 147–156.

Dandara, C., Masimirembwa, C. M., Magimba, A., Sayi, J., Kaaya, S., Sommers, D. K., … Hasler, J. A. (2001). Genetic polymorphism of CYP2D6 and CYP2C19 in east- and southern African populations including psychiatric patients. *European Journal of Clinical Pharmacology*, **57**(1), 11–17.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, **27**(15), 2156–2158.

da Rocha, J. E. B., Othman, H., Botha, G., Cottino, L., Twesigomwe, D., Ahmed, S., et al. (2021). The extent and impact of variation in ADME genes in sub-Saharan African populations. *Frontiers in Pharmacology*, **12**, 366.

da Rocha, J. E. B., Othman, H., Tiemessen, C. T., Botha, G., Ramsay, M., and Masimirembwa, C. (2021). *G6PD* variant distribution in sub-Saharan Africa and potential risks of using chloroquine/hydroxychloroquine based treatments for COVID-19. *The Pharmacogenomics Journal*, 1-8, Epub ahead of print.

Das, A. K., Goswami, S., Lee, K., and Park, S.-J. (2019). A hybrid and scalable error correction algorithm for indel and substitution errors of long reads. *BMC Genomics*, **20**(Suppl 11), 948.

Desta, Z., El-Boraie, A., Gong, L., Somogyi, A. A., Lauschke, V. M., Dandara, C., et al. (2021). PharmVar GeneFocus: *CYP2B6*. *Clinical Pharmacology and Therapeutics*, **110**(1), 82–97.

Desta, Z., Gammal, R. S., Gong, L., Whirl-Carrillo, M., Gaur, A. H., Sukasem, C., et al. (2019). Clinical pharmacogenetics implementation consortium (CPIC) guideline for *CYP2B6* and efavirenz-containing antiretroviral therapy. *Clinical Pharmacology and Therapeutics*, **106**(4), 726–733.

Desta, Z., Metzger, I. F., Thong, N., Lu, J. B. L., Callaghan, J. T., Skaar, T. C., et al. (2016). Inhibition of Cytochrome P450 2B6 activity by voriconazole profiled using efavirenz disposition in healthy volunteers. *Antimicrobial Agents and Chemotherapy*, **60**(11), 6813–6822.

Dhoro, M., Zvada, S., Ngara, B., Nhachi, C., Kadzirange, G., Chonzi, P., and Masimirembwa, C. (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 and Toxicology*, **16**, 4.

Di, Y. M., Chow, V. D.-W., Yang, L.-P., and Zhou, S.-F. (2009). Structure, function, regulation and polymorphism of human cytochrome P450 2A6. *Current Drug Metabolism*, **10**(7), 754–780.

di Lulio, J., Fayet, A., Arab-Alameddine, M., Rotger, M., Lubomirov, R., Cavassini, M., et al. (2009). *in vivo* analysis of efavirenz metabolism in individuals with impaired CYP2A6 function. *Pharmacogenetics and Genomics*, **19**(4), 300–309.

Diouf, B., Crews, K. R., Lew, G., Pei, D., Cheng, C., Bao, J., et al. (2015). Association of an inherited genetic variant with vincristine-related peripheral neuropathy in children with acute lymphoblastic leukemia. *JAMA*, **313**(8), 815–823.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**(4), 316–319.

Dodgen, T. M., Eloff, A., Mataboge, C., Roos, L. J., van Staden, W. C., and Pepper, M. S. (2015). Risperidone-

associated adverse drug reactions and CYP2D6 polymorphisms in a South African cohort. *Applied & translational genomics*, **5**, 40–46.

Dodgen, T. M., Hochfeld, W. E., Fickl, H., Asfaha, S. M., Durandt, C., Rheeder, P., et al. (2013). Introduction of the AmpliChip CYP450 test to a South African cohort: a platform comparative prospective cohort study. *BMC medical genetics*, **14**, 20.

Dodgen, T. M., Labuschagne, C. J., van Schalkwyk, A., Steffens, F. E., Gaedigk, A., Cromarty, A. D., ... Pepper, M. S. (2016). Pharmacogenetic comparison of CYP2D6 predictive and measured phenotypes in a South African cohort. *The Pharmacogenomics Journal*, **16**(6), 566–572.

Dohm, J. C., Peters, P., Stralis-Pavese, N., and Himmelbauer, H. (2020). Benchmarking of long-read correction methods. *NAR Genomics and Bioinformatics*, **2**(2).

Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., ... Eberle, M. A. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*, **35**(22), 4754–4756.

Drögemöller, B. I., Plummer, M., Korkie, L., Agenbag, G., Dunaiski, A., Niehaus, D., et al. (2013). Characterization of the genetic variation present in *CYP3A4* in three South African populations. *Frontiers in Genetics*, **4**, 17.

Drögemöller, B. I., Wright, G. E. B., Niehaus, D. J. H., Emsley, R., and Warnich, L. (2013). Next-generation sequencing of pharmacogenes: a critical analysis focusing on schizophrenia treatment. *Pharmacogenetics and Genomics*, **23**(12), 666–674.

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., et al. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, **49**(11), 1654–1660.

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., ... Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, **10**(1), 5402.

Eisenstein, M. (2015). Startups use short-read data to expand long-read sequencing market. *Nature Biotechnology*, **33**(5), 433–435.

El-Boraie, A., Chenoweth, M. J., Pouget, J. G., Benowitz, N. L., Fukunaga, K., Mushiroda, T., ... Tyndale, R. F. (2021). Transferability of Ancestry-specific and cross-ancestry CYP2A6 activity genetic risk scores in African and European populations. *Clinical Pharmacology and Therapeutics*, **110**(4), 975–985.

Fernandez, C., Smith, C., Yang, W., Lorier, R., Crews, K., Kornegay, N., et al. (2012). Concordance of DMET plus genotypes with orthogonal genotyping methods. *Clinical pharmacology and therapeutics*, **92**(3), 360–365.

Flockhart, D. A., Thacker, D., McDonald, C., and Desta, Z. (2021). The Flockhart Cytochrome P450 Drug-Drug Interaction Table. Division of Clinical Pharmacology, Indiana University School of Medicine (Updated 2021). Accessed June 08, 2021, from `https://drug-interactions.medicine.iu.edu`.

Fukami, T., Nakajima, M., Yamanaka, H., Fukushima, Y., McLeod, H. L., and Yokoi, T. (2007). A novel duplication type of *CYP2A6* gene in African-American population. *Drug Metabolism and Disposition*, **35**(4), 515–520.

Fukami, T., Nakajima, M., Yoshida, R., Tsuchiya, Y., Fujiki, Y., Katoh, M., et al. (2004). A novel polymorphism of human *CYP2A6* gene *CYP2A6*17* has an amino acid substitution (v365m) that decreases enzymatic activity *in*

vitro and *in vivo*. *Clinical Pharmacology and Therapeutics*, **76**(6), 519–527.

Gaedigk, A. (2013). Complexities of *CYP2D6* gene analysis and interpretation. *International Review of Psychiatry*, **25**(5), 534–553.

Gaedigk, A., Blum, M., Gaedigk, R., Eichelbaum, M., and Meyer, U. A. (1991). Deletion of the entire cytochrome P450 *CYP2D6* gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. *American Journal of Human Genetics*, **48**(5), 943–950.

Gaedigk, A., Casey, S. T., Whirl-Carrillo, M., Miller, N. A., and Klein, T. E. (2021). Pharmacogene variation consortium: A global resource and repository for pharmacogene variation. *Clinical Pharmacology and Therapeutics*.

Gaedigk, A., and Coetsee, C. (2008). The *CYP2D6* gene locus in South African Coloureds: unique allele distributions, novel alleles and gene arrangements. *European Journal of Clinical Pharmacology*, **64**(5), 467–475.

Gaedigk, A., Dinh, J. C., Jeong, H., Prasad, B., and Leeder, J. S. (2018). Ten years' experience with the *CYP2D6* activity score: A perspective on future investigations to improve clinical predictions for precision therapeutics. *Journal of Personalized Medicine*, **8**(2), E15.

Gaedigk, A., Fuhr, U., Johnson, C., Bérard, L. A., Bradford, D., and Leeder, J. S. (2010). *CYP2D7-2D6* hybrid tandems: identification of novel *CYP2D6* duplication arrangements and implications for phenotype prediction. *Pharmacogenomics*, **11**(1), 43–53.

Gaedigk, A., Ingelman-Sundberg, M., Miller, N. A., Leeder, J. S., Whirl-Carrillo, M., Klein, T. E., et al. (2018). The pharmacogene variation (PharmVar) consortium: Incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clinical Pharmacology and Therapeutics*, **103**(3), 399–401.

Gaedigk, A., Ndjountché, L., Divakaran, K., Bradford, D. L., Zineh, I., Oberlander, T. F., et al. (2007). Cytochrome P450 2D6 (*CYP2D6*) gene locus heterogeneity: characterization of gene duplication events. *Clinical Pharmacology and Therapeutics*, **81**(2), 242–251.

Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Klein, T., and Leeder, J. S. (2017). Prediction of CYP2D6 phenotype from genotype across world populations. *Genetics in Medicine*, **19**(1), 69–76.

Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Twist, G. P., Klein, T. E., and Miller, N. A. (2019). The evolution of PharmVar. *Clinical Pharmacology and Therapeutics*, **105**(1), 29–32.

Gaedigk, A., Simon, S. D., Pearce, R. E., Bradford, L. D., Kennedy, M. J., and Leeder, J. S. (2008). The *CYP2D6* activity score: translating genotype information into a qualitative measure of phenotype. *Clinical Pharmacology and Therapeutics*, **83**(2), 234–242.

Gaedigk, A., Turner, A., Everts, R. E., Scott, S. A., Aggarwal, P., Broeckel, U., et al. (2019). Characterization of reference materials for genetic testing of *CYP2D6* alleles: A GeT-RM collaborative project. *The Journal of molecular diagnostics*, **21**(6), 1034–1052.

Gaedigk, A., Twist, G. P., and Leeder, J. S. (2012). *CYP2D6*, *SULT1A1* and *UGT2B17* copy number variation: quantitative detection by multiplex PCR. *Pharmacogenomics*, **13**(1), 91–111.

Gaedigk, A. et al. (2021). *CYP2D6*149.001* allele definition. Retrieved from `https://www.pharmvar.org/gene/CYP2D6`.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, **36**(9), 875–879.

GATK team. (2021). Genome Analysis Toolkit tutorial: (How to) Map reads to a reference with alternate contigs like GRCH38. Accessed August 18, 2021, from `https://gatk.broadinstitute.org/hc/en-us/articles/360037498992--How-to-Map-reads-to-a-reference-with-alternate-contigs-like-GRCH38`.

Gelston, E. A., Coller, J. K., Lopatko, O. V., James, H. M., Schmidt, H., White, J. M., and Somogyi, A. A. (2012). Methadone inhibits CYP2D6 and UGT2B7/2B4 *in vivo*: a study using codeine in methadone- and buprenorphine-maintained subjects. *British Journal of Clinical Pharmacology*, **73**(5), 786–794.

Goetz, M. P., Sangkuhl, K., Guchelaar, H. J., Schwab, M., Province, M., Whirl-Carrillo, M., ... Klein, T. E. (2018). Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6 and tamoxifen therapy. *Clinical Pharmacology and Therapeutics*, **103**(5), 770–777.

Gomez, F., Hirbo, J., and Tishkoff, S. A. (2014). Genetic variation and adaptation in Africa: Implications for human evolution and disease. *Cold Spring Harbor Perspectives in Biology*, **6**(7), a008524.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**(6), 333–351.

Gordon, A., Fulton, R. S., Qin, X., Mardis, E. R., Nickerson, D. A., and Scherer, S. (2016). PGRNseq: A targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenetics and genomics*, **26**(4), 161–168.

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

Gulilat, M., Lamb, T., Teft, W. A., Wang, J., Dron, J. S., Robinson, J. F., et al. (2019). Targeted next generation sequencing as a tool for precision medicine. *BMC Medical Genomics*, **12**(1), 81.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, **517**(7534), 327–332.

Gutiérrez Rico, E. M., Kikuchi, A., Saito, T., Kumondai, M., Hishinuma, E., Kaneko, A., ... Hiratsuka, M. (2020). CYP2D6 genotyping analysis and functional characterization of novel allelic variants in a Ni-Vanuatu and Kenyan population by assessing dextromethorphan O-demethylation activity. *Drug Metabolism and Pharmacokinetics*, **35**(1), 89–101.

Hakooz, N., and Hamdan, I. (2007). Effects of dietary broccoli on human *in vivo* caffeine metabolism: a pilot study on a group of Jordanian volunteers. *Current Drug Metabolism*, **8**(1), 9–15.

Harewood, L., Kishore, K., Eldridge, M. D., Wingett, S., Pearson, D., Schoenfelder, S., et al. (2017). Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biology*, **18**(1), 125.

Hayat, M., Kerr, R., Bentley, A. R., Rotimi, C. N., Raal, F. J., and Ramsay, M. (2020). Genetic associations between serum low LDL-cholesterol levels and variants in *LDLR*, *APOB*, *PCSK9* and *LDLRAP1* in African populations.

PLOS ONE, **15**(2), e0229098.

Heather, J. M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, **107**(1), 1–8.

Heim, M. H., and Meyer, U. A. (1992). Evolution of a highly polymorphic human cytochrome P450 gene cluster: CYP2D6. *Genomics*, **14**(1), 49–58.

Hesse, L. M., He, P., Krishnaswamy, S., Hao, Q., Hogan, K., von Moltke, L. L., … Court, M. H. (2004). Pharmacogenetic determinants of interindividual variability in bupropion hydroxylation by cytochrome P450 2B6 in human liver microsomes. *Pharmacogenetics*, **14**(4), 225–238.

Higashi, E., Fukami, T., Itoh, M., Kyo, S., Inoue, M., Yokoi, T., and Nakajima, M. (2007). Human CYP2A6 is induced by estrogen via estrogen receptor. *Drug Metabolism and Disposition*, **35**(10).

Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, **21**(3), 171–189.

Hoffman, S. M., Nelson, D. R., and Keeney, D. S. (2001). Organization, structure and evolution of the *CYP2* gene cluster on human chromosome 19. *Pharmacogenetics*, **11**(8), 687–698.

Hofmann, M. H., Blievernicht, J. K., Klein, K., Saussele, T., Schaeffeler, E., Schwab, M., and Zanger, U. M. (2008). Aberrant splicing caused by single nucleotide polymorphism c.516G>T [Q172H], a marker of *CYP2B6*6*, is responsible for decreased expression and activity of CYP2B6 in liver. *The Journal of Pharmacology and Experimental Therapeutics*, **325**(1), 284–292.

Hosono, N., Kato, M., Kiyotani, K., Mushiroda, T., Takata, S., Sato, H., et al. (2009). *CYP2D6* genotyping for functional-gene dosage analysis by allele copy number detection. *Clinical Chemistry*, **55**(8), 1546–1554.

Hovelson, D. H., Xue, Z., Zawistowski, M., Ehm, M. G., Harris, E. C., Stocker, S. L., et al. (2017). Characterization of ADME gene variation in 21 populations by exome sequencing. *Pharmacogenetics and Genomics*, **27**(3), 89–100.

Hu, J., and Ng, P. C. (2013). SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLOS ONE*, **8**(10), e77940.

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., et al. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications*, **6**, 8111.

Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., et al. (2017). A reference human genome dataset of the BGISEQ-500 sequencer. *GigaScience*, **6**(5), 1–9.

Hukkanen, J., Jacob, P., and Benowitz, N. L. (2006). Effect of grapefruit juice on cytochrome P450 2A6 and nicotine renal clearance. *Clinical Pharmacology and Therapeutics*, **80**(5), 522–530.

Institute for Health Metrics and Evaluation. (2019). GBD-compare Health Data [Visualisation Hub]. Accessed July 15, 2021, from `https://vizhub.healthdata.org/gbd-compare/`.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Isah, A. O., Pal, S. N., Olsson, S., Dodoo, A., and Bencheikh, R. S. (2012). Specific features of medicines safety and pharmacovigilance in Africa. *Therapeutic Advances in Drug Safety*, **3**(1), 25–34.

Itoh, M., Nakajima, M., Higashi, E., Yoshida, R., Nagata, K., Yamazoe, Y., and Yokoi, T. (2006). Induction of human CYP2A6 is mediated by the pregnane X receptor with peroxisome proliferator-activated receptor-gamma coactivator 1alpha. *The Journal of Pharmacology and Experimental Therapeutics*, **319**(2), 693–702.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, **36**(4), 338–345.

Jeon, S. A., Park, J. L., Kim, J.-H., Kim, J. H., Kim, Y. S., Kim, J. C., and Kim, S.-Y. (2019). Comparison of the MGISEQ-2000 and illumina HiSeq 4000 sequencing platforms for RNA sequencing. *Genomics and Informatics*, **17**(3), e32.

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., et al. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, **21**(1), 189.

Johansson, I., Lundqvist, E., Bertilsson, L., Dahl, M. L., Sjöqvist, F., and Ingelman-Sundberg, M. (1993). Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proceedings of the National Academy of Sciences of the United States of America*, **90**(24), 11825–11829.

Johnson, J. A., Caudle, K. E., Gong, L., Whirl-Carrillo, M., Stein, C. M., Scott, S. A., et al. (2017). Clinical pharmacogenetics implementation consortium (CPIC) guideline for pharmacogenetics-guided warfarin dosing: 2017 update. *Clinical Pharmacology and Therapeutics*, **102**(3), 397–404.

Kamp, M., Krause, A., and Ramsay, M. (2021). Has translational genomics come of age in Africa? *Human Molecular Genetics*, Epub ahead of print.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2018). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**(7809), 434–443.

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**(8), 907–915.

Kimura, S., Umeno, M., Skoda, R. C., Meyer, U. A., and Gonzalez, F. J. (1989). The human debrisoquine 4-hydroxylase (*CYP2D*) locus: sequence and identification of the polymorphic *CYP2D6* gene, a related gene, and a pseudogene. *American Journal of Human Genetics*, **45**(6), 889–904.

Klein, H. G., Flegel, W. A., and Natanson, C. (2015). Red blood cell transfusion: Precision vs imprecision medicine. *JAMA*, **314**(15), 1557–1558.

Klein, K., Lang, T., Saussele, T., Barbosa-Sicard, E., Schunck, W. H., Eichelbaum, M., ... Zanger, U. M. (2005). Genetic variability of CYP2B6 in populations of African and Asian origin: allele frequencies, novel functional variants, and possible implications for anti-HIV therapy with efavirenz. *Pharmacogenetics and Genomics*, **15**(12), 861–873.

Klein, K., Tremmel, R., Winter, S., Fehr, S., Battke, F., Scheurenbrand, T., et al. (2019). A new panel-based next-generation sequencing method for ADME genes reveals novel associations of common and rare variants with expression in a human liver cohort. *Frontiers in Genetics*, **10**, 7.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, **155**(1), 27–38.

Koenigs, L. L., Peter, R. M., Thompson, S. J., Rettie, A. E., and Trager, W. F. (1997). Mechanism-based inactivation

of human liver cytochrome P450 2A6 by 8-methoxypsoralen. *Drug Metabolism and Disposition*, **25**(12), 1407–1415.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, **30**(7), 693–700.

Kramer, W. E., Walker, D. L., O'Kane, D. J., Mrazek, D. A., Fisher, P. K., Dukek, B. A., … Black, J. L. (2009). *CYP2D6*: novel genomic structures and alleles. *Pharmacogenetics and Genomics*, **19**(10), 813–822.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, **4**(7), 1073–1081.

Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLOS ONE*, **12**(5), e0177459.

Lang, T., Klein, K., Richter, T., Zibat, A., Kerb, R., Eichelbaum, M., … Zanger, U. M. (2004). Multiple novel nonsynonymous CYP2B6 gene polymorphisms in Caucasians: demonstration of phenotypic null alleles. *The Journal of Pharmacology and Experimental Therapeutics*, **311**(1), 34–43.

Lee, S.-B., Wheeler, M. M., Patterson, K., McGee, S., Dalton, R., Woodahl, E. L., et al. (2019). Stargazer: a software tool for calling star alleles from next-generation sequencing data using *CYP2D6* as a model. *Genetics in Medicine*, **21**(2), 361–372.

Lee, S.-B., Wheeler, M. M., Thummel, K. E., Nickerson, D. A., et al. (2019). Calling star alleles with stargazer in 28 pharmacogenes with whole genome sequences. *Clinical Pharmacology and Therapeutics*, **106**(6), 1328–1337.

Lee, S. S., Cha, E.-Y., Jung, H.-J., Shon, J.-H., Kim, E.-Y., Yeo, C.-W., and Shin, J.-G. (2008). Genetic polymorphism of hepatocyte nuclear factor-4alpha influences human cytochrome P450 2D6 activity. *Hepatology*, **48**(2), 635–645.

Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K. Y., McCaffrey, J., et al. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications*, **10**(1), 1025.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, **21**(1), 265.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

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

Liang, L., Morar, N., Dixon, A. L., Lathrop, G. M., Abecasis, G. R., Moffatt, M. F., and Cookson, W. O. C. (2013). A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Research*, **23**(4), 716–726.

Liau, Y., Maggo, S., Miller, A. L., Pearson, J. F., Kennedy, M. A., and Cree, S. L. (2019). Nanopore sequencing of the pharmacogene *CYP2D6* allows simultaneous haplotyping and detection of duplications. *Pharmacogenomics*,

20(14), 1033–1047.

Liu, Y., Cheng, J., Siejka-Zielińska, P., Weldon, C., Roberts, H., Lopopolo, M., et al. (2020). Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biology*, **21**(1), 54.

LoTempio, J., Délot, E., and Vilain, E. (2021). *Benchmarking long-read genome sequence alignment tools for human genomics applications.*

M., H.-Y. F., Leinonen, R., Cochrane, G., and Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, **21**(5), 734–740.

MacDougall, J. M., Fandrick, K., Zhang, X., Serafin, S. V., and Cashman, J. R. (2003). Inhibition of human liver microsomal (S)-nicotine oxidation by (-)-menthol and analogues. *Chemical Research in Toxicology*, **16**(8), 988–993.

Maimbo, M., Kiyotani, K., Mushiroda, T., Masimirembwa, C., and Nakamura, Y. (2012). *CYP2B6* genotype is a strong predictor of systemic exposure to efavirenz in HIV-infected Zimbabweans. *European Journal of Clinical Pharmacology*, **68**(3), 267–271.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., ... others (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, **538**(7624), 201–206.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, **51**(4), 584–591.

Martiniano, R., Garrison, E., Jones, E. R., Manica, A., and Durbin, R. (2020). Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biology*, **21**(1), 250.

Martis, S., Mei, H., Vijzelaar, R., Edelmann, L., Desnick, R. J., and Scott, S. A. (2013). Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. *The Pharmacogenomics Journal*, **13**(6), 558–566.

Masimirembwa, C., Persson, I., Bertilsson, L., Hasler, J., and Ingelman-Sundberg, M. (1996). A novel mutant variant of the *CYP2D6* gene (*CYP2D6*17*) common in a black African population: association with diminished debrisoquine hydroxylase activity. *British Journal of Clinical Pharmacology*, **42**(6), 713–719.

Matimba, A., Del-Favero, J., Van Broeckhoven, C., and Masimirembwa, C. (2009). Novel variants of major drug-metabolising enzyme genes in diverse African populations and their predicted functional effects. *Human Genomics*, **3**(2), 169–190.

Mauleekoonphairoj, J., Chamnanphon, M., Khongphatthanayothin, A., Sutjaporn, B., Wandee, P., Poovorawan, Y., et al. (2020). Phenotype prediction and characterization of 25 pharmacogenes in Thais from whole genome sequencing for clinical implementation. *Scientific Reports*, **10**(1), 18969.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, **48**(10), 1279–1283.

McGinn, S., and Gut, I. G. (2013). DNA sequencing - spanning the generations. *New Biotechnology*, **30**(4), 366–372.

McInnes, G., Dalton, R., Sangkuhl, K., Whirl-Carrillo, M., Lee, S.-B., Tsao, P. S., et al. (2020). Transfer learning

enables prediction of *CYP2D6* haplotype function. *PLOS Computational Biology*, **16**(11), e1008399.

McInnes, G., Lavertu, A., Sangkuhl, K., Klein, T. E., Whirl-Carrillo, M., and Altman, R. B. (2021). Pharmacogenetics at scale: An analysis of the UK biobank. *Clinical Pharmacology and Therapeutics*, **109**(6), 1528–1537.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, **19**(9), 1527–1541.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., ... Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, **17**(1), 122.

Mehlotra, R. K., Ziats, M. N., Bockarie, M. J., and Zimmerman, P. A. (2006). Prevalence of *CYP2B6* alleles in malaria-endemic populations of West Africa and Papua New Guinea. *European Journal of Clinical Pharmacology*, **62**(4), 267-275.

Messina, E. S., and Tyndale, E. M., R. F. Sellers. (1997). A major role for CYP2A6 in nicotine C-oxidation by human liver microsomes. *The Journal of Pharmacology and Experimental Therapeutics*, **282**, 1608–1614.

Meyer, U. A., Zanger, U. M., and Schwab, M. (2013). Omics and drug response. *Annual Review of Pharmacology and Toxicology*, **53**, 475–502.

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, **585**(7823), 79–84.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, **16**(9), 1182–1190.

Morozova, O., and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**(5), 255–264.

Mugusi, S., Ngaimisi, E., Janabi, M., Minzi, O., Bakari, M., Riedel, K.-D., et al. (2012). Liver enzyme abnormalities and associated risk factors in HIV patients on efavirenz-based HAART with or without tuberculosis co-infection in tanzania. *PLOS ONE*, **7**(7), e40180.

Mukonzo, J. K., Nanzigu, S., Waako, P., Ogwal-Okeng, J., Gustafson, L. L., and Aklillu, E. (2014). CYP2B6 genotype, but not rifampicin-based anti-TB cotreatments, explains variability in long-term efavirenz plasma exposure. *Pharmacogenomics*, **15**(11), 1423–1435.

Mutawi, T. M., Zedan, M. M., Yahya, R. S., M., Z. M., El-Sawi, M. R., and A., G. (2021). Genetic variability of *CYP2D6*, *CYP3A4* and *CYP3A5* among the egyptian population. *Pharmacogenomics*, **22**(6), 323–334.

Mwenifumbo, J. C., Al Koudsi, N., Ho, M. K., Zhou, Q., Hoffmann, E. B., Sellers, E. M., and Tyndale, R. F. (2008). Novel and established *CYP2A6* alleles impair in vivo nicotine metabolism in a population of Black African descent. *Human Mutation*, **29**(5), 679–688.

Mwenifumbo, J. C., Zhou, Q., Benowitz, N. L., Sellers, E. M., and Tyndale, R. F. (2010). New *CYP2A6* gene deletion and conversion variants in a population of black African descent. *Pharmacogenomics*, **11**(2), 189–198.

Nakajima, M., Fukami, T., Yamanaka, H., Higashi, E., Sakai, H., Yoshida, R., et al. (2006). Comprehensive

evaluation of variability in nicotine metabolism and *CYP2A6* polymorphic alleles in four ethnic populations. *Clinical Pharmacology and Therapeutics*, **80**(3), 282–297.

Nebert, D. W. (1999). Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist? *Clinical Genetics*, **56**(4), 247–258.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, **541**(7637), 302–310.

Nofziger, C., and Paulmichl, M. (2018). Accurately genotyping *CYP2D6*: not for the faint of heart. *Pharmacogenomics*, **19**(13), 999–1002.

Nofziger, C., Turner, A. J., Sangkuhl, K., Whirl-Carrillo, M., Agúndez, J. A. G., Black, J. L., et al. (2020). PharmVar GeneFocus: *CYP2D6*. *Clinical Pharmacology and Therapeutics*, **107**(1), 154–170.

Numanagić, I., Malikić, S., Ford, M., Qin, X., Toji, L., Radovich, M., et al. (2018). Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nature Communications*, **9**(1), 828.

Numanagić, I., Malikić, S., Pratt, V. M., Skaar, T. C., Flockhart, D. A., and Sahinalp, S. C. (2015). Cypiripi: exact genotyping of *CYP2D6* using high-throughput sequencing data. *Bioinformatics*, **31**(12), i27–34.

Nyakutira, C., Röshammar, D., Chigutsa, E., Chonzi, P., Ashton, M., Nhachi, C., and Masimirembwa, C. (2008). High prevalence of the *CYP2B6* 516G–>T(**6*) variant and effect on the population pharmacokinetics of efavirenz in HIV/AIDS outpatients in Zimbabwe. *European Journal of Clinical Pharmacology*, **64**(4), 357–365.

Ogburn, E. T., Jones, D. R., Masters, A. R., Xu, C., Guo, Y., and Desta, Z. (2010). Efavirenz primary and secondary metabolism *in vitro* and *in vivo*: identification of novel metabolic pathways and cytochrome P450 2A6 as the principal catalyst of efavirenz 7-hydroxylation. *Drug Metabolism and Disposition*, **38**(7), 1218-1229.

Oscarson, M., McLellan, R. A., Asp, V., Ledesma, M., Bernal Ruiz, M. L., Sinues, B., et al. (2002). Characterization of a novel *CYP2A7/CYP2A6* hybrid allele (*CYP2A6*12*) that causes reduced CYP2A6 activity. *Human Mutation*, **20**(4), 275–283.

Oscarson, M., McLellan, R. A., Gullstén, H., Yue, Q. Y., Lang, M. A., Bernal, M. L., et al. (1999). Characterisation and PCR-based detection of a *CYP2A6* gene deletion found at a high frequency in a Chinese population. *FEBS letters*, **448**(1), 105–110.

Othman, H., da Rocha, J. E. B., and Hazelhurst, S. (2021). Single nucleotide polymorphism induces divergent dynamic patterns in *CYP3A5*: A microsecond scale biomolecular simulation of variants identified in sub-Saharan African populations. *International Journal of Molecular Sciences*, **22**(15), 7786.

Ovenden, E. S., Drögemöller, B. I., van der Merwe, L., Chiliza, B., Asmal, L., Emsley, R. A., and Warnich, L. (2017). Fine-mapping of antipsychotic response genome-wide association studies reveals novel regulatory mechanisms. *Pharmacogenomics*, **18**(2), 105–120.

Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G. W., and Schönhuth, A. (2015). WhatsHap: Weighted haplotype assembly for future-Generation sequencing reads. *Journal of Computational Biology*, **22**(6), 498-509.

Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2019). BulkVis: a graphical viewer for Oxford Nanopore bulk FAST5 files. *Bioinformatics*, **35**(13), 2193–2198.

Peko, S. M., Gueye, N. S. G., Vouvoungui, C., Koukouikila-Koussounda, F., Kobawila, S. C., Nderu, D., et al. (2019). Cytochrome P450 *CYP2B6*6* distribution among Congolese individuals with HIV, tuberculosis and malaria infection. *International Journal of Infectious Diseases*, **82**, 111–116.

Pereira, L., Mutesa, L., Tindana, P., and Ramsay, M. (2021). African genetic diversity and adaptation inform a precision medicine agenda. *Nature Reviews Genetics*, **22**(5), 284–306.

Pernaute-Lau, L., Adegnika, A. A., Zhou, Y., Zinsou, J. F., Gil, J. P., Krishna, S., ... Velavan, T. P. (2021). Pharmacogene sequencing of a Gabonese population with severe Plasmodium falciparum malaria reveals multiple novel variants with putative relevance for antimalarial treatment. *Antimicrobial Agents and Chemotherapy*, **65**(7), e0027521.

Petros, Z., Lee, M. T. M., Takahashi, A., Zhang, Y., Yimer, G., Habtewold, A., et al. (2017). Genome-wide association and replication study of hepatotoxicity induced by antiretrovirals alone or with concomitant anti-tuberculosis drugs. *OMICS*, **21**(4), 207–216.

PharmGKB. (2021). What is Pharmacogenomics? Accessed May 07, 2021, from `https://www.pharmgkb.org/whatIsPharmacogenomics`.

PharmGKB and CPIC *CYP2D6* reference materials. (2021). Retrieved August 16, 2021, from `https://www.pharmgkb.org/page/cyp2d6RefMaterials`.

PharmGKB gene-specific information tables. (2021). Retrieved August 08, 2021, from `https://www.pharmgkb.org/page/pgxGeneRef`.

PharmVar structural variation for *CYP2B6*. (2021). Accessed July 28, 2021, from `https://a.storyblok.com/f/70677/x/090e5f7e50/cyp2b6_variation_v1-2.pdf`.

PharmVar structural variation for *CYP2D6*. (2021). Accessed July 17, 2021, from `https://a.storyblok.com/f/70677/x/7073c64512/cyp2d6_structural-variation_v2-1.pdf`.

Piliguian, M., Zhu, A. Z. X., Zhou, Q., Benowitz, N. L., Ahluwalia, J. S., Sanderson Cox, L., and Tyndale, R. F. (2014). Novel *CYP2A6* variants identified in African Americans are associated with slow nicotine metabolism *in vitro* and *in vivo*. *Pharmacogenetics and Genomics*, **24**(2), 118–128.

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., ... DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, **36**(10), 983–987.

Pratt, V. M., Everts, R. E., Aggarwal, P., Beyer, B. N., Broeckel, U., Epstein-Baak, R., et al. (2016). Characterization of 137 genomic DNA reference materials for 28 pharmacogenetic genes: A GeT-RM collaborative project. *The Journal of Molecular Diagnostics*, **18**(1), 109–123.

Qiao, W., Wang, J., Pullman, B. S., Chen, R., Yang, Y., and Scott, S. A. (2017). The CYP2D6 VCF translator. *The Pharmacogenomics Journal*, **17**(4), 301–303.

Qiao, W., Yang, Y., Sebra, R., Mendiratta, G., Gaedigk, A., Desnick, R. J., and Scott, S. A. (2016). Long-read single molecule real-time full gene sequencing of cytochrome P450-2D6. *Human Mutation*, **37**(3), 315–323.

Radloff, R., Gras, A., Zanger, U. M., Masquelier, C., Arumugam, K., Karasi, J.-C., et al. (2013). Novel CYP2B6 enzyme variants in a rwandese population: functional characterization and assessment of in silico prediction tools. *Human Mutation*, **34**(5), 725–734.

Radouani, F., Zass, L., Hamdi, Y., da Rocha, J., Sallam, R., Abdelhak, S., et al. (2020). A review of clinical pharmacogenetics studies in African populations. *Personalized Medicine*, **17**(2), 155–170.

Rae, J. M., Johnson, M. D., Lippman, M. E., and Flockhart, D. A. (2001). Rifampin is a selective, pleiotropic inducer of drug metabolism genes in human hepatocytes: studies with cDNA and oligonucleotide expression arrays. *The Journal of Pharmacology and Experimental Therapeutics*, **299**(3), 849–857.

Rajman, I., Knapp, L., Morgan, T., and Masimirembwa, C. (2017). African genetic diversity: Implications for cytochrome P450-mediated drug metabolism and drug development. *EBioMedicine*, **17**, 67–74.

Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., et al. (2019). Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, **51**(2), 354–362.

Ramsay, M., Crowther, N., Tambo, E., Agongo, G., Baloyi, V., Dikotope, S., 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**, e20.

Ramsey, L. B., Johnson, S. G., Caudle, K. E., Haidar, C. E., Voora, D., Wilke, R. A., et al. (2014). The clinical pharmacogenetics implementation consortium guideline for *SLCO1B1* and simvastatin-induced myopathy: 2014 update. *Clinical Pharmacology and Therapeutics*, **96**(4), 423–428.

Rang, F. J., Kloosterman, W. P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, **19**(1), 90.

Rao, Y., Hoffmann, E., Zia, M., Bodin, L., Zeman, M., Sellers, E. M., and Tyndale, R. F. (2000). Duplications and defects in the *CYP2A6* gene: identification, genotyping, and *in vivo* effects on smoking. *Molecular Pharmacology*, **58**(4), 747–755.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature*, **444**(7118), 444–454.

Relling, M., and Klein, T. (2011). CPIC: Clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clinical Pharmacology and Therapeutics*, **89**(3), 464–467.

Relling, M. V., Klein, T. E., Gammal, R. S., Whirl-Carrillo, M., Hoffman, J. M., and Caudle, K. E. (2020). The clinical pharmacogenetics implementation consortium: 10 years later. *Clinical Pharmacology and Therapeutics*, **107**(1), 171–175.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, **47**(D1), D886–D894.

Richter, T., Mürdter, T. E., Heinkele, G., Pleiss, J., Tatzel, S., Schwab, M., et al. (2004). Potent mechanism-based inhibition of human CYP2B6 by clopidogrel and ticlopidine. *The Journal of Pharmacology and Experimental Therapeutics*, **308**(1), 189–197.

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology*, **14**(6), 405.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, **29**(1), 24-26.

Rotger, M., Saumoy, M., Zhang, K., Flepp, M., Sahli, R., Decosterd, L., et al. (2007). Partial deletion of *CYP2B6* owing to unequal crossover with *CYP2B7*. *Pharmacogenetics and Genomics*, **17**(10), 885–890.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**(7356), 348–352.

Roy, S., and Motsinger-Reif, A. (2013). Evaluation of calling algorithms for array-CGH. *Frontiers in Genetics*, **4**, 217.

Röhrich, C. R., Drögemöller, B. I., Ikediobi, O., van der Merwe, L., Grobbelaar, N., Wright, G. E., et al. (2016). *CYP2B6*6* and *CYP2B6*18* predict long-term efavirenz exposure measured in hair samples in HIV-positive South African women. *AIDS Research and Human Retroviruses*, **32**(6), 529–538.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., et al. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**(4839), 487–491.

Saitoh, A., Fletcher, C. V., Brundage, R., Alvero, C., Fenton, T., Hsia, K., and Spector, S. A. (2007). Efavirenz pharmacokinetics in hiv-1-infected children are associated with *CYP2B6*-g516t polymorphism. *Journal of Acquired Immune Deficiency Syndrome*, **45**(3), 280–285.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463–5467.

Sangkuhl, K., Whirl-Carrillo, M., Whaley, R. M., Woon, M., Lavertu, A., Altman, R. B., et al. (2020). Pharmacogenomics clinical annotation tool (PharmCAT). *Clinical Pharmacology and Therapeutics*, **107**(1), 203–210.

Sariya, S., Lee, J. H., Mayeux, R., Vardarajan, B. N., Reyes-Dumeyer, D., Manly, J. J., et al. (2019). Rare variants imputation in admixed populations: Comparison across reference panels and bioinformatics tools. *Frontiers in Genetics*, **10**, 239.

Scott, S. A. (2011). Personalizing medicine with clinical pharmacogenetics. *Genetics in Medicine*, **13**(12), 987–995.

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews. Genetics*, **19**(6), 329–346.

Sengupta, D., Choudhury, A., Fortes-Lima, C., Aron, S., Whitelaw, G., Bostoen, K., et al. (2021). Genetic substructure and complex demographic history of South African Bantu speakers. *Nature Communications*, **12**(1), 2080.

Shah, R. R., and Smith, R. L. (2015). Addressing phenoconversion: the achilles' heel of personalized medicine. *British Journal of Clinical Pharmacology*, **79**(2), 222–240.

Shelton, J. M., Coleman, M. C., Herndon, N., Lu, N., Lam, E. T., Anantharaman, T., et al. (2015). Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics*, **16**(1), 734.

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, **51**(1), 30–35.

Sherry, S. T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other

classes of minor genetic variation. *Genome Research*, **9**(8), 677–679.

Shimada, T., Yamazaki, H., Mimura, M., Inui, Y., and Guengerich, F. P. (1994). Interindividual variations in human liver cytochrome P-450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: studies with liver microsomes of 30 Japanese and 30 Caucasians. *The Journal of Pharmacology and Experimental Therapeutics*, **270**(1), 414–423.

Sirén, J., Garrison, E., Novak, A. M., Paten, B., and Durbin, R. (2020). Haplotype-aware graph indexes. *Bioinformatics*, **36**(2), 400–407.

Siu, E. C. K., and Tyndale, R. F. (2008). Selegiline is a mechanism-based inactivator of CYP2A6 inhibiting nicotine metabolism in humans and mice. *The Journal of Pharmacology and Experimental Therapeutics*, **324**(3), 992–999.

Soeria-Atmadja, S., Österberg, E., Gustafsson, L. L., Dahl, M. L., Eriksen, J., Rubin, J., and Navér, L. (2012). Genetic variants in CYP2B6 and CYP2A6 explain interindividual variation in efavirenz plasma concentrations of HIV-infected children with diverse ethnic origin. *PlOS ONE*, **12**(9), e0181316.

Steen, V. M., Andreassen, O. A., Daly, A. K., Tefre, T., Børresen, A. L., Idle, J. R., and Gulbrandsen, A. K. (1995). Detection of the poor metabolizer-associated *CYP2D6*(D) gene deletion allele by long-PCR technology. *Pharmacogenetics*, **5**(4), 215–223.

Strasser, A. A., Benowitz, N. L., Pinto, A. G., Tang, K. Z., Hecht, S. S., Carmella, S. G., et al. (2011). Nicotine metabolite ratio predicts smoking topography and carcinogen biomarker level. *Cancer Epidemiology, Biomarkers and Prevention*, **20**(2), 234–238.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

Swart, M., Skelton, M., Ren, Y., Smith, P., Takuva, S., and Dandara, C. (2013). High predictive value of *CYP2B6* SNPs for steady-state plasma efavirenz levels in South African HIV/AIDS patients. *Pharmacogenetics and Genomics*, **23**(8), 415–427.

Tanner, J.-A., and Tyndale, R. F. (2017). Variation in CYP2A6 activity and personalized medicine. *Journal of Personalized Medicine*, **7**(4), E18.

Thomford, N. E., Awortwe, C., Dzobo, K., Adu, F., Chopera, D., Wonkam, A., et al. (2016). Inhibition of CYP2B6 by medicinal plant extracts: Implication for use of efavirenz and nevirapine-based highly active anti-retroviral therapy (HAART) in resource-limited settings. *Molecules*, **21**(2), 211.

Thomford, N. E., Dzobo, K., Adu, F., Chirikure, S., Wonkam, A., and Dandara, C. (2018). Bush mint (hyptis suaveolens) and spreading hogweed (boerhavia diffusa) medicinal plant extracts differentially affect activities of CYP1A2, CYP2D6 and CYP3A4 enzymes. *Journal of Ethnopharmacology*, **211**, 58–69.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., et al. (2009). The genetic structure and history of Africans and African Americans. *Science*, **324**(5930), 1035–1044.

Tourdot, R. W., Brunette, G. J., Pinto, R. A., and Zhang, C.-Z. (2021). Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biology*, **22**(1), 139.

Turner, A. J., Aggarwal, P., Boone, E. C., Haidar, C.-E., Relling, M. V., Derezinski, A. D., et al. (2021). Identification of *CYP2D6* haplotypes that interfere with commonly used assays for copy number variation characterization. *The Journal of Molecular Diagnostics*, **23**(5), 577–588.

Twist, G. P., Gaedigk, A., Miller, N. A., Farrow, E. G., Willig, L. K., Dinwiddie, D. L., et al. (2016). Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, *CYP2D6*, from whole-genome sequences. *npj Genomic Medicine*, **1**(1), 1–10.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **43**, 11.10.1–11.10.33.

van der Merwe, N., Bouwens, C. S., Pienaar, R., van der Merwe, L., Yako, Y. Y., Geiger, D. H., and Kotze, M. J. (2012). CYP2D6 genotyping and use of antidepressants in breast cancer patients: test development for clinical application. *Metabolic Brain Disease*, **27**(3), 319–326.

Vergara, C., Parker, M. M., Franco, L., Cho, M. H., Valencia-Duarte, A. V., Beaty, T. H., and Duggal, P. (2018). Genotype imputation performance of three reference panels using African ancestry individuals. *Human Genetics*, **137**(4), 281–292.

von Moltke, L. L., Greenblatt, D. J., Grassi, J. M., Granda, B. W., Venkatakrishnan, K., Schmider, J., ... Shader, R. I. (1998). Multiple human cytochromes contribute to biotransformation of dextromethorphan in-vitro: role of CYP2C9, CYP2C19, CYP2D6, and CYP3A. *The Journal of Pharmacy and Pharmacology*, **50**(9), 997–1004.

Vos, T., Lim, S. S., Abbafati, K. M., C. Abbas, Abbasi, M., Abbasifard, M., et al. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*, **396**(10258), 1204–1222.

Wang, D., Poi, M. J., Sun, X., Gaedigk, A., Leeder, J. S., and Sadee, W. (2014). Common *CYP2D6* polymorphisms affecting alternative splicing and transcription: long-range haplotypes with two regulatory variants modulate CYP2D6 activity. *Human Molecular Genetics*, **23**(1), 268–278.

Wang, H., and Tompkins, L. M. (2008). CYP2B6: New insights into a historically overlooked cytochrome P450 isozyme. *Current Drug Metabolism*, **9**(7), 598–610.

Wang, J., Pitarque, M., and Ingelman-Sundberg, M. (2006). 3'-UTR polymorphism in the human *CYP2A6* gene affects mRNA stability and enzyme expression. *Biochemical and Biophysical Research Communications*, **340**(2), 491–497.

Ware, J. J., Tanner, J.-A., Taylor, A. E., Bin, Z., Haycock, P., Bowden, J., et al. (2017). Does coffee consumption impact on heaviness of smoking? *Addiction*, **112**(10), 1842–1853.

Wassenaar, C. A., Zhou, Q., and Tyndale, R. F. (2016). *CYP2A6* genotyping methods and strategies using real-time and endpoint PCR platforms. *Pharmacogenomics*, **17**(2), 147–162.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, **37**(10), 1155–1162.

Wennerholm, A., Dandara, C., Sayi, J., Svensson, J.-O., Abdi, Y. A., Ingelman-Sundberg, M., et al. (2002). The African-specific *CYP2D6*17* allele encodes an enzyme with changed substrate specificity. *Clinical Pharmacology and Therapeutics*, **71**(1), 77–88.

Wennerholm, A., Johansson, I., Hidestrand, M., Bertilsson, L., Gustafsson, L. L., and Ingelman-Sundberg, M. (2001). Characterization of the *CYP2D6*29* allele commonly present in a black Tanzanian population causing

reduced catalytic activity. *Pharmacogenetics*, **11**(5), 417–427.

Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F., Whaley, R., and Klein, T. E. (2021). An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*, Epub ahead of print.

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., et al. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*, **92**(4), 414–417.

World Health Organisation. (2020). Noncommunicable diseases progress monitor. Retrieved from `https://www.who.int/publications/i/item/ncd-progress-monitor-2020`.

Wright, G. E. B., Carleton, B., Hayden, M. R., and Ross, C. J. D. (2018). The global spectrum of protein-coding pharmacogenomic diversity. *The Pharmacogenomics Journal*, **18**(1), 187–195.

Wright, G. E. B., Niehaus, D. J. H., Drögemöller, B. I., Koen, L., Gaedigk, A., and Warnich, L. (2010). Elucidation of *CYP2D6* genetic diversity in a unique African population: implications for the future application of pharmacogenetics in the Xhosa population. *Annals of Human Genetics*, **74**(4), 340–350.

Wu, T. D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**(7), 873–881.

Yang, Y., Botton, M. R., Scott, E. R., and Scott, S. A. (2017). Sequencing the *CYP2D6* gene: from variant allele discovery to clinical pharmacogenetic testing. *Pharmacogenomics*, **18**(7), 673–685.

Yimer, G., Gry, M., Amogne, W., Makonnen, E., Habtewold, A., Petros, Z., et al. (2014). Evaluation of patterns of liver toxicity in patients on antiretroviral and anti-tuberculosis drugs: A prospective four arm observational study in Ethiopian patients. *PLOS ONE*, **9**(4), e94271.

Yoshida, R., Nakajima, M., Watanabe, Y., Kwon, J.-T., and Yokoi, T. (2002). Genetic polymorphisms in human *CYP2A6* gene causing impaired nicotine metabolism. *British Journal of Clinical Pharmacology*, **54**(5), 511–517.

Yuan, Y., Chung, C. Y.-L., and Chan, T.-F. (2020). Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal*, **18**, 2051–2062.

Zanger, U. M., Klein, K., Saussele, T., Blievernicht, J., Hofmann, M. H., and Schwab, M. (2007). Polymorphic *CYP2B6*: molecular mechanisms and emerging clinical significance. *Pharmacogenomics*, **8**(7), 743–759.

Zanger, U. M., and Schwab, M. (2013). Cytochrome p450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology and Therapeutics*, **138**(1), 103–141.

Zeng, L., Chen, Y., Wang, Y., Yu, L.-R., Knox, B., Chen, J., et al. (2017). MicroRNA hsa-miR-370-3p suppresses the expression and induction of CYP2D6 by facilitating mRNA degradation. *Biochemical Pharmacology*, **140**, 139–149.

Zhang, H., De, T., Zhong, Y., and Perera, M. A. (2019). The advantages and challenges of diversity in pharmacogenomics: Can minority populations bring us closer to implementation? *Clinical Pharmacology and Therapeutics*, **106**(2), 338–349.

Zhang, W., Kilicarslan, T., Tyndale, R. F., and Sellers, E. M. (2001). Evaluation of methoxsalen, tranylcypromine, and tryptamine as specific and selective CYP2A6 inhibitors *in vitro*. *Drug Metabolism and Disposition*, **29**(6),

897–902.

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, **34**(3), 303–311.

Zhou, S.-F. (2009). Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I. *Clinical Pharmacokinetics*, **48**(11), 689–723.

Zhou, Y., Ingelman-Sundberg, M., and Lauschke, V. M. (2017). Worldwide distribution of cytochrome P450 alleles: A meta-analysis of population-scale sequencing projects. *Clinical Pharmacology and Therapeutics*, **102**(4), 688–700.

Zhou, Y., Mkrtchian, S., Kumondai, M., Hiratsuka, M., and Lauschke, V. M. (2019). An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *The Pharmacogenomics Journal*, **19**, 115–126.

Zhu, A. Z., Cox, L. S., Nollen, N., Faseru, B., Okuyemi, K. S., Ahluwalia, J. S., … Tyndale, R. F. (2012). CYP2B6 and bupropion's smoking-cessation pharmacology: the role of hydroxybupropion. *British Journal of Clinical Pharmacology*, **92**(6), 771–777.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, **3**, 160025.

Zook, J. M., McDaniel, J., Olson, N. D., Wagner, J., Parikh, H., Heaton, H., et al. (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, **37**(5), 561–566.

Zubiaur, P., Saiz-Rodríguez, M., Ochoa, D., Belmonte, C., Román, M., Mejía, G., … Abad-Santos, F. (2020). Influence of CYP2B6 activity score on the pharmacokinetics and safety of single dose efavirenz in healthy volunteers. *The Pharmacogenomics Journal*, **20**(2), 235–245.

Appendices

Appendix A: Authors' agreement

Declaration: Student's contribution to articles and agreement of co-authors

I, David Twesigomwe, student number 1989491, declare that this Thesis is my own work and that I contributed adequately towards research findings published in the articles stated below which are included in my Thesis.

Signature of student ..... **Date** 06/12/2021

Name of primary supervisor Prof Scott Hazelhurst

Signature of primary supervisor Scott Hazelhurst (Dec 17, 2021 11:17 GMT+2) **Date** Dec 17, 2021

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of the article by the student as part of his Thesis.

Article 1

Title: A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: A focus on *CYP2D6* genotyping.

Journal name, year, volume and page numbers: *npj Genomic Medicine*, 2020, 5, 30. https://doi.org/10.1038/s41525-020-0135-2.

Authors	Name	Signature	Date
1st author	David Twesigomwe		06/12/2021
2nd author	Galen E. B. Wright	Galen Wright (Dec 7, 2021 18:15 CST)	07/12/2021
3rd author	Britt I. Drögemöller	Britt Drogemoller (Dec 7, 2021 18:11 CST)	07/12/2021
4th author	Jorge da Rocha		07/12/2021
5th author	Zané Lombard		17/12/2021
6th author	Scott Hazelhurst	Scott Hazelhurst (Dec 17, 2021 11:17 GMT+2)	17/12/2021

Comments by primary supervisor:

David initiated this work and did the bulk of the work

Article 2

Title: StellarPGx: A Nextflow pipeline for calling star alleles in cytochrome P450 genes.

Journal name, year, volume and page numbers: *Clinical Pharmacology and Therapeutics*, 2021, 110 (3), 741-749. https://doi.org/10.1002/cpt.2173.

Authors	Name	Signature	Date
1st author	David Twesigomwe		06/12/2021
2nd author	Britt I. Drögemöller	 Britt Drogemoller (Dec 7, 2021 18:11 CST)	07/12/2021
3rd author	Galen E. B. Wright	 Galen Wright (Dec 7, 2021 18:15 CST)	07/12/2021
4th author	Azra Siddiqui	 Azra Siddiqui (Dec 8, 2021 09:14 GMT+2)	08/12/2021
5th author	Jorge da Rocha		07/12/2021
6th author	Zané Lombard		17/12/2021
7th author	Scott Hazelhurst	 Scott Hazelhurst (Dec 17, 2021 11:17 GMT+2)	18/12/2021

Comments by primary supervisor:

David initiated this work and was the primary driver

Appendix B: Ethics clearance certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Mr D Twesigomwe

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M200993

NAME: (Principal Investigator)	Mr D Twesigomwe
DEPARTMENT:	School of Pathology Department of Human Genetics Medical School University
PROJECT TITLE:	Characterization of pharmacogene allelic varation in African populations and development of a novel diplotype calling algorithm
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M170880
SUPERVISOR:	Professors S Hazelhurst and Z Lombard
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2020/10/06

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

2020/10/1
Date

Appendix C: *CYP2D6* haplotypes used for simulation of test samples used in the allele caller benchmarking study.

Table C7.1: Set 1 and 2 SNV-defined haplotypes used to generate simulated datasets (Chapter 3). Some of the haplotypes contain the benign intron 1 conversion.

#	Haplotype (*allele/suballele)	PharmVar Level of Evidence	Notes on core variant(s)	Highest population frequency [a]
1	*CYP2D6*1.001*	Definitive	None	AF ≥ 1 %
2	*CYP2D6*1.002*	Definitive		
3	*CYP2D6*1.003*	Limited		
4	*CYP2D6*1.004*	Limited		
5	*CYP2D6*1.005*	Definitive		
6	*CYP2D6*1.007*	Definitive		
7	*CYP2D6*1.008*	Definitive		
8	*CYP2D6*1.014*	Definitive		
9	*CYP2D6*1.019*	Definitive		
10	*CYP2D6*1.025*	Definitive		
11	*CYP2D6*2.001*	Definitive	SNP core variant(s)	AF ≥ 1 %
12	*CYP2D6*2.002*	Limited		
13	*CYP2D6*2.003*	Limited		
14	*CYP2D6*2.004*	Limited		
15	*CYP2D6*2.007*	Limited		
16	*CYP2D6*2.010*	Limited		
17	*CYP2D6*2.011*	Limited		
18	*CYP2D6*2.013*	Definitive		
19	*CYP2D6*2.014*	Definitive		
20	*CYP2D6*2.017*	Definitive		
21	*CYP2D6*3.001*	Definitive	Indel core variant(s)	AF ≥ 1 %
22	*CYP2D6*3.002*	Limited		
23	*CYP2D6*4.001*	Definitive	SNP core variant(s)	AF ≥ 1 %
24	*CYP2D6*4.002*	Limited		
25	*CYP2D6*4.003*	Limited		
26	*CYP2D6*4.004*	Definitive		
27	*CYP2D6*4.007*	Limited		
28	*CYP2D6*4.009*	Limited		
29	*CYP2D6*4.012*	Definitive		
30	*CYP2D6*4.014*	Definitive		
31	*CYP2D6*4.021*	Definitive		
32	*CYP2D6*4.026*	Definitive		
33	*CYP2D6*6.001*	Limited	Indel core variant(s)	0.1 < AF < 1%
34	*CYP2D6*6.002*	Limited		
35	*CYP2D6*6.003*	Limited		
36	*CYP2D6*6.004*	Limited		
37	*CYP2D6*6.005*	Definitive		

38	*CYP2D6*6.006*	Definitive		
39	*CYP2D6*7.001*	Definitive	SNP core variant(s)	AF < 0.1%
40	*CYP2D6*8.001*	Limited	SNP core variant(s)	AF < 0.1%
41	*CYP2D6*9.001*	Definitive	Indel core variant(s)	AF ≥ 1 %
42	*CYP2D6*9.002*	Definitive	Indel core variant(s)	
43	*CYP2D6*10.001*	Limited	SNP core variant(s)	AF ≥ 1 %
44	*CYP2D6*10.002*	Definitive		
45	*CYP2D6*10.004*	Definitive		
46	*CYP2D6*11.001*	Definitive	SNP core variant(s)	AF < 0.1%
47	*CYP2D6*12.001*	Limited	SNP core variant(s)	AF ≥ 1 %
48	*CYP2D6*12.002*	Definitive		
49	*CYP2D6*14.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
50	*CYP2D6*15.001*	Definitive	Indel core variant(s)	0.1 < AF < 1%
51	*CYP2D6*15.002*	Definitive		
52	*CYP2D6*15.003*	Definitive		
53	*CYP2D6*17.001*	Definitive	SNP core variant(s)	AF ≥ 1 %
54	*CYP2D6*17.002*	Definitive		
55	*CYP2D6*17.003*	Definitive		
56	*CYP2D6*18.001*	Limited	Indel core variant(s)	AF < 0.1%
57	*CYP2D6*19.001*	Limited	Indel core variant(s)	0.1 < AF < 1%
58	*CYP2D6*20.001*	Limited	Indel core variant(s)	AF < 0.1%
59	*CYP2D6*21.001*	Limited	Indel core variant(s)	0.1 < AF < 1%
60	*CYP2D6*21.002*	Definitive		
61	*CYP2D6*22.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
62	*CYP2D6*23.001*	Limited	SNP core variant(s)	AF < 0.1%
63	*CYP2D6*24.001*	Limited	SNP core variant(s)	AF < 0.1%
64	*CYP2D6*25.001*	Limited	SNP core variant(s)	AF < 0.1%
65	*CYP2D6*26.001*	Limited	SNP core variant(s)	AF < 0.1%
66	*CYP2D6*27.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
67	*CYP2D6*28.001*	Definitive	SNP core variant(s)	AF < 0.1%
68	*CYP2D6*28.002*	Definitive		
69	*CYP2D6*29.001*	Definitive	SNP core variant(s)	AF > 1%
70	*CYP2D6*30.001*	Limited	Indel core variant(s)	AF < 0.1%
71	*CYP2D6*31.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
72	*CYP2D6*33.001*	Definitive	SNP core variant(s)	AF > 1%
73	*CYP2D6*34.001*	Limited	SNP core variant(s)	AF > 1%
74	*CYP2D6*35.001*	Definitive	SNP core variant(s)	AF > 1%
75	*CYP2D6*35.002*	Definitive		
76	*CYP2D6*35.003*	Definitive		
77	*CYP2D6*35.004*	Definitive		
78	*CYP2D6*35.006*	Definitive		

79	*CYP2D6*35.007*	Definitive		
80	*CYP2D6*37.001*	Limited	SNP core variant(s)	AF < 0.1%
81	*CYP2D6*38.001*	Limited	Indel core variant(s)	AF < 0.1%
82	*CYP2D6*39.001*	Definitive	SNP core variant(s)	AF > 1%
83	*CYP2D6*40.001*	Definitive	Indel core variant(s)	AF > 1%
84	*CYP2D6*41.001*	Definitive	SNP core variant(s)	AF > 1%
85	*CYP2D6*41.002*	Definitive		
86	*CYP2D6*41.003*	Definitive		
87	*CYP2D6*41.004*	Definitive		
88	*CYP2D6*41.005*	Definitive		
89	*CYP2D6*42.001*	Definitive	Indel core variant(s)	0.1 < AF < 1%
90	*CYP2D6*43.001*	Definitive	SNP core variant(s)	AF > 1%
91	*CYP2D6*43.002*	Definitive		
92	*CYP2D6*44.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
93	*CYP2D6*45.001*	Definitive	SNP core variant(s)	AF > 1%
94	*CYP2D6*45.002*	Definitive		
95	*CYP2D6*46.001*	Definitive	SNP core variant(s)	AF > 1%
96	*CYP2D6*46.002*	Definitive		
97	*CYP2D6*46.003*	Definitive		
98	*CYP2D6*47.001*	Definitive	SNP core variant(s)	AF < 0.1%
99	*CYP2D6*48.001*	Definitive	SNP core variant(s)	AF > 1%
100	*CYP2D6*49.001*	Definitive	SNP core variant(s)	AF > 1%
101	*CYP2D6*50.001*	Definitive	SNP core variant(s)	AF < 0.1%
102	*CYP2D6*51.001*	Definitive	SNP core variant(s)	AF < 0.1%
103	*CYP2D6*52.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
104	*CYP2D6*52.002*	Definitive		
105	*CYP2D6*53.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
106	*CYP2D6*54.001*	Limited	SNP core variant(s)	AF < 0.1%
107	*CYP2D6*55.001*	Limited	SNP core variant(s)	AF < 0.1%
108	*CYP2D6*56.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
109	*CYP2D6*56.002*	Definitive		
110	*CYP2D6*56.003*	Definitive		
111	*CYP2D6*58.001*	Definitive	Indel core variant(s)	0.1 < AF < 1%
112	*CYP2D6*59.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
113	*CYP2D6*60.001*	Limited	Indel core variant(s)	AF < 0.1%
114	*CYP2D6*62.001*	Definitive	SNP core variant(s)	AF < 0.1%
115	*CYP2D6*64.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
116	*CYP2D6*65.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
117	*CYP2D6*69.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
118	*CYP2D6*70.001*	Moderate	SNP core variant(s)	AF < 0.1%
119	*CYP2D6*71.001*	Moderate	SNP core variant(s)	0.1 < AF < 1%

120	*CYP2D6*71.002*	Definitive		
121	*CYP2D6*71.003*	Definitive		
122	*CYP2D6*72.001*	Limited	SNP core variant(s)	AF < 0.1%
123	*CYP2D6*73.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
124	*CYP2D6*74.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
125	*CYP2D6*75.001*	Moderate	SNP core variant(s)	AF < 0.1%
126	*CYP2D6*81.001*	Moderate	SNP core variant(s)	AF < 0.1%
127	*CYP2D6*84.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
128	*CYP2D6*84.002*	Definitive		
129	*CYP2D6*85.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
130	*CYP2D6*86.001*	Definitive	SNP core variant(s)	0.1 < AF < 1%
131	*CYP2D6*87.001*	Limited	SNP core variant(s)	AF < 0.1%
132	*CYP2D6*88.001*	Limited	SNP core variant(s)	AF < 0.1%
133	*CYP2D6*89.001*	Limited	SNP core variant(s)	AF < 0.1%
134	*CYP2D6*90.001*	Definitive	SNP core variant(s)	AF < 0.1%
135	*CYP2D6*91.001*	Limited	SNP core variant(s)	AF < 0.1%
136	*CYP2D6*92.001*	Limited	SNP core variant(s)	AF < 0.1%
137	*CYP2D6*93.001*	Limited	SNP core variant(s)	AF < 0.1%
138	*CYP2D6*94.001*	Limited	SNP core variant(s)	AF < 0.1%
139	*CYP2D6*94.002*	Limited		
140	*CYP2D6*95.001*	Limited	SNP core variant(s)	AF < 0.1%
141	*CYP2D6*96.001*	Limited	SNP core variant(s)	AF < 0.1%
142	*CYP2D6*97.001*	Limited	SNP core variant(s)	AF < 0.1%
143	*CYP2D6*98.001*	Limited	SNP core variant(s)	AF < 0.1%
144	*CYP2D6*100.001*	Definitive	Indel core variant(s)	0.1 < AF < 1%
145	*CYP2D6*101.001*	Definitive	Indel core variant(s)	0.1 < AF < 1%
146	*CYP2D6*102.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
147	*CYP2D6*103.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
148	*CYP2D6*104.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
149	*CYP2D6*105.001*	Limited	SNP core variant(s)	0.1 < AF < 1%
150	*CYP2D6*107.001*	Limited	SNP core variant(s)	AF < 0.1%
151	*CYP2D6*108.001*	Definitive	SNP core variant(s)	AF < 0.1%
152	*CYP2D6*109.001*	Definitive	Indel core variant(s)	AF < 0.1%
153	*CYP2D6*106.001*	Definitive	SNP core variant(s)	AF < 0.1%
154	*CYP2D6*106.002*	Definitive		

Detailed variant descriptions can be found at `https://www.pharmvar.org/gene/CYP2D6`.

[a] See `https://www.pharmgkb.org/page/cyp2d6RefMaterials` for more information on population frequencies.

Table C7.2: *CYP2D6* structural variants used to generate the simulated Set 3 dataset in Chapter 3.

#	Haplotype (*allele/arrangement)	Notes	Gene and/or pseudogene involved
1	**4.013* (with REP6)	Exon 9 conversion	*CYP2D6 & CYP2D7*
2	**35.002*	Conversion upstream of exon 1	*CYP2D6 & CYP2D7*
3	**36.001* (with REP6)	Exon 9 conversion	*CYP2D6 & CYP2D7*
4	**36.002* (with REP6)	Exon 9 conversion	*CYP2D6 & CYP2D7*
5	**57*	Exon 9 conversion	*CYP2D6 & CYP2D7*
6	**82*	Exon 2 conversion	*CYP2D6 & CYP2D7*
7	**83.001*	Exon 9 conversion	*CYP2D6 & CYP2D7*
8	**83.002*	Exon 9 conversion	*CYP2D6 & CYP2D7*
9	**83.003*	Exon 9 conversion	*CYP2D6 & CYP2D7*
10	**1x2*	gene duplication	*CYP2D6*
11	**2x2*	gene duplication	*CYP2D6*
12	**3x2*	gene duplication	*CYP2D6*
13	**4x2*	gene duplication	*CYP2D6*
14	**6x2*	gene duplication	*CYP2D6*
15	**9x2*	gene duplication	*CYP2D6*
16	**10x2*	gene duplication	*CYP2D6*
17	**17x2*	gene duplication	*CYP2D6*
18	**29x2*	gene duplication	*CYP2D6*
19	**35x2*	gene duplication	*CYP2D6*
20	**41x2*	gene duplication	*CYP2D6*
21	**43x2*	gene duplication	*CYP2D6*
22	**45x2*	gene duplication	*CYP2D6*
23	**1xN*	gene multiplication	*CYP2D6*
24	**2xN*	gene multiplication	*CYP2D6*
25	**4xN*	gene multiplication	*CYP2D6*
26	**10xN*	gene multiplication	*CYP2D6*
27	**41xN*	gene multiplication	*CYP2D6*
28	**4.013*	2D6-2D7 hybrid	*CYP2D6 & CYP2D7*
29	**10.003*	2D6-2D7 hybrid	*CYP2D6 & CYP2D7*
30	**13 (*16, *66, *67, *76, *77, *78, *79)*	2D7-2D6 hybrid(s)	*CYP2D6 & CYP2D7*
31	**36*	2D6-2D7 hybrid	*CYP2D6 & CYP2D7*
32	**61*	2D6-2D7 hybrid	*CYP2D6 & CYP2D7*
33	**63*	2D6-2D7 hybrid	*CYP2D6 & CYP2D7*
34	**68*	2D6-2D7 hybrid	*CYP2D6 & CYP2D7*
35	**1x2+*83*	tandem rearrangement	*CYP2D6 & CYP2D7*
36	**1+*90*	tandem rearrangement	*CYP2D6*
37	**4.013+*4*	tandem rearrangement	*CYP2D6 & CYP2D7*
38	**4.013xN +*4*	tandem rearrangement	*CYP2D6 & CYP2D7*

39	**4.013 + *4xN*	tandem rearrangement	*CYP2D6 & CYP2D7*
40	**13 +*1*	tandem rearrangement	*CYP2D6 & CYP2D7*
41	**13+*1x2*	tandem rearrangement	*CYP2D6 & CYP2D7*
42	**13+*2*	tandem rearrangement	*CYP2D6 & CYP2D7*
43	**13+*68+*4*	tandem rearrangement	*CYP2D6 & CYP2D7*
44	**17* (REP7) *+ *17*	tandem rearrangement	*CYP2D6 & CYP2D7*
45	**36* (REP7) *+ *36*	tandem rearrangement	*CYP2D6 & CYP2D7*
46	**36x2 +*10*	tandem rearrangement	*CYP2D6 & CYP2D7*
47	**36+*10*	tandem rearrangement	*CYP2D6 & CYP2D7*
48	**36+*10x2*	tandem rearrangement	*CYP2D6 & CYP2D7*
49	**57+*10*	tandem rearrangement	*CYP2D6 & CYP2D7*
50	**68+*4*	tandem rearrangement	*CYP2D6 & CYP2D7*

Appendix D: Supplementary materials for Chapter 4.

Sample StellarPGx output files

1. Sample with previously described alleles

Consider the screenshot below showing an example StellarPGx output file for a **simulated** sample with the *CYP2D6*17/*29* diplotype. The input was a whole genome sequence Binary Alignment Map (BAM) file with short read data aligned to GRCh38.

```
-----------------------------------------

CYP2D6 Star Allele Calling with StellarPGx

-----------------------------------------

Initially computed CN = 2


Core variants:
42126611~C>G~1/1;42127608~C>T~0/1;42127941~G>A~1/1;42129132~C>T~0/1;42129770~G>A~0/1


Candidate alleles:
['17.v1_29.v1']


Result:
*17/*29


Activity score:
1.0


Metaboliser status:
Intermediate metaboliser (IM)
```

*Sample StellarPGx output file for a simulated sample with the CYP2D6*17/*29 diplotype*

Note:

'17.v1 and 29.v1' is a unique notation that StellarPGx uses to report **candidate** alleles. Each of these possible candidate haplotypes is defined in the StellarPGx database. For the final result, StellarPGx reports alleles using the normal star (*) nomenclature. The importance of the <allele>.v1, <allele>.v2, … notation is clearer for alleles such as *CYP2D6*4, CYP2D6*3, and CYP2D6*6*. For instance, there are a number of *CYP2D6*4* suballeles whose core variants form part of the definition of other haplotypes. This representation could be useful to users who would like more information about the allele calls from a sample. For example, consider *CYP2D6*6* core variants in the table below:

Haplotype	Candidate allele	Core variants (hg38)
*CYP2D6*6.001, CYP2D6*6.002, CYP2D6*6.004, CYP2D6*6.005, CYP2D6*6.006*	6.v1	42129083~CA>C
*CYP2D6*6.003*	6.v2	42126611~C>G;42129083~CA>C

Whether a sample has *CYP2D6*6.003 (*6.v2) or one of the other *CYP2D6*6* suballeles (denoted using 6.v1 when using StellarPGx) can make a difference in star allele calling depending on the allele present on the other chromosome.

2. Sample with a potential novel allele

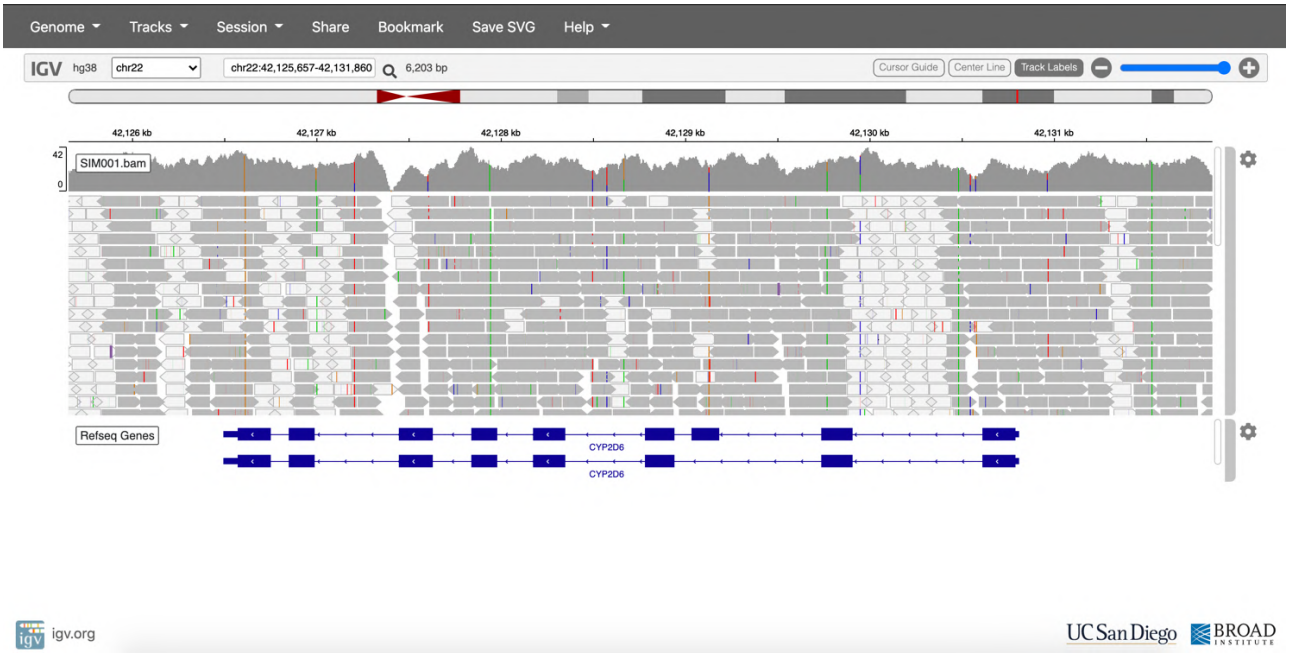
Consider the screenshot below showing a sample StellarPGx output for a **simulated** sample with a novel allele on a *CYP2D6*17* background. Notice that the rs61745683 (42127512C>T, *122) variant (heterozygous in this case) has been detected. Currently, there is no known haplotype comprising the *CYP2D6*17* and *CYP2D6*122* core variant(s) in linkage disequilibrium (https://www.pharmvar.org/gene/CYP2D6), thus prompting the novel allele flag in the StellarPGx output. The input was a whole genome sequence Binary Alignment Map (BAM) file with short read data aligned to GRCh38.

```
-----------------------------------------

CYP2D6 Star Allele Calling with StellarPGx

-----------------------------------------

Initially computed CN = 2

Sample core variants:
42126611~C>G~1/1;42127512~C>T~0/1;42127941~G>A~1/1;42129770~G>A~1/1

Result:
Possible novel allele or suballele present: interpret with caution;
Experimental validation and expert review through PharmVar is recommended

Likely background alleles:
[*17/*17]

Activity score:
Indeterminate

Metaboliser status:
Indeterminate
```

*Sample StellarPGx output for a simulated sample with a novel allele on a CYP2D6*17 background.*

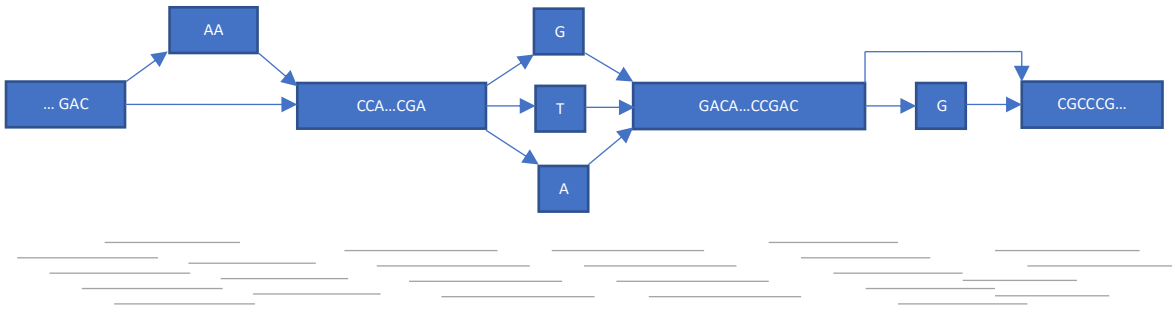
Illustration of StellarPGx's combinatorial diplotype assignment approach

For this description, we use simulated sample SIM001 (*CYP2D6*17.001/*29.001*) as an example. Allele definitions are obtained from the PharmVar (Gaedigk et al., 2018) database.



Integrative Genomics Viewer (*Robinson et al., 2011) plot showing read alignment in the CYP2D6 locus for sample SIM001.*

Step 1: *CYP2D6* reads are realigned to a pangenome structure (i.e. a variant-aware reference genome) using GraphTyper (Eggertsson et al., 2017)



Step 1: Model showing a variant-aware reference genome (i.e. a pangenome)

Step 2: GraphTyper outputs a Variant Call Format (VCF) file with high-confidence *CYP2D6* Small Nucloetide Variant (SNV) calls for the sample.

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SIM001
chr22   42126611        chr22:42126611:SG       C       G       255     PASS    ABHet=-1;ABHom=1;AC=2;AF=1
;AN=2;CR=0;LOGF=0.9756;MQ=58;MaxAAS=21;MaxAASR=1;NHet=0;NHomAlt=1;NHomRef=0;PASS_AC=2;PASS_AN=2;PASS_ratio
=1;QD=25.5;RefLen=1;SB=0.4762;SBAlt=0.4762;SBF=0,10;SBF1=0,0;SBF2=0,10;SBR=0,11;SBR1=0,0;SBR2=0,11;SeqDept
h=21;VarType=SG   GT:AD:MD:DP:GQ:PL       1/1:0,21:0:21:63:255,63,0
chr22   42127608        chr22:42127608:SG       C       T       39      PASS    ABHet=0.2727;ABHom=-1;AC=1
;AF=0.5;AN=2;CR=0;LOGF=1.35e-11;MQ=47;MaxAAS=3;MaxAASR=0.2727;NHet=1;NHomAlt=0;NHomRef=0;PASS_AC=0;PASS_AN
=0;PASS_ratio=0;QD=13;RefLen=1;SB=0.5455;SBAlt=0.6667;SBF=4,2;SBF1=0,0;SBF2=4,2;SBR=4,1;SBR1=0,0;SBR2=4,1;
SeqDepth=11;VarType=SG    GT:AD:MD:DP:GQ:PL       0/1:8,3:0:11:39:39,0,144
chr22   42127941        chr22:42127941:SG       G       A       255     PASS    ABHet=-1;ABHom=1;AC=2;AF=1
;AN=2;CR=0;LOGF=0.974;MQ=54;MaxAAS=20;MaxAASR=1;NHet=0;NHomAlt=1;NHomRef=0;PASS_AC=2;PASS_AN=2;PASS_ratio=
1;QD=25.5;RefLen=1;SB=0.5;SBAlt=0.5;SBF=0,10;SBF1=0,0;SBF2=0,10;SBR=0,10;SBR1=0,0;SBR2=0,10;SeqDepth=20;Va
rType=SG  GT:AD:MD:DP:GQ:PL       1/1:0,20:0:20:60:255,60,0
chr22   42129130        chr22:42129130:SG       C       G       255     PASS    ABHet=-1;ABHom=1;AC=2;AF=1
;AN=2;CR=0;LOGF=0.9716;MQ=43;MaxAAS=14;MaxAASR=1;NHet=0;NHomAlt=1;NHomRef=0;PASS_AC=2;PASS_AN=2;PASS_ratio
=1;QD=25.5;RefLen=1;SB=0.4286;SBAlt=0.4286;SBF=0,6;SBF1=0,0;SBF2=0,6;SBR=0,8;SBR1=0,0;SBR2=0,8;SeqDepth=14
;VarType=SG       GT:AD:MD:DP:GQ:PL       1/1:0,14:0:14:96:255,96,0
chr22   42129132        chr22:42129132:SG       C       T       69      PASS    ABHet=0.3571;ABHom=-1;AC=1
;AF=0.5;AN=2;CR=0;LOGF=6.096e-11;MQ=43;MaxAAS=5;MaxAASR=0.3571;NHet=1;NHomAlt=0;NHomRef=0;PASS_AC=1;PASS_A
N=2;PASS_ratio=1;QD=13.8;RefLen=1;SB=0.4286;SBAlt=0.6;SBF=3,3;SBF1=0,0;SBF2=3,3;SBR=6,2;SBR1=0,0;SBR2=6,2;
SeqDepth=14;VarType=SG    GT:AD:MD:DP:GQ:PL       0/1:9,5:0:14:69:69,0,157
chr22   42129770        chr22:42129770:SG       G       A       255     PASS    ABHet=0.6522;ABHom=-1;AC=1
;AF=0.5;AN=2;CR=0;LOGF=1.255e-09;MQ=60;MaxAAS=15;MaxAASR=0.6522;NHet=1;NHomAlt=0;NHomRef=0;PASS_AC=1;PASS_
AN=2;PASS_ratio=1;QD=25.5;RefLen=1;SB=0.4783;SBAlt=0.5333;SBF=3,8;SBF1=0,0;SBF2=3,8;SBR=5,7;SBR1=0,0;SBR2=
5,7;SeqDepth=23;VarType=SG        GT:AD:MD:DP:GQ:PL       0/1:8,15:0:23:99:255,0,114
```

Step 2 output: GraphTyper VCF output for simulated test sample SIM001

Step 3: StellarPGx identifies possible allele defining variants from the sample VCF. For SIM001, the core variants include: 42126611~C>G~1/1; 42127608~C>T~0/1; 42127941~G>A~1/1; 42129132~C>T~0/1; 42129770~G>A~0/1.

Step 4: StellarPGx queries the sample's variant combination against a database of all *CYP2D6* core variant combinations, thus deducing *CYP2D6*17* and *29 as candidate alleles.

```
15.v2_97.v1     42126697~G>T~0/1;42130655~A>AA~0/1;42130715~C>T~0/1      42126697~G>T;42130469~C>T;42130
15.v2_98.v1     42126611~C>G~0/1;42126681~G>C~0/1;42127941~G>A~0/1;42130655~A>AA~0/1;42130715~C>T~0/1
15.v2_99.v1     42126611~C>G~0/1;42129827~C>G~0/1;42130655~A>AA~0/1;42130692~G>A~0/1;42130715~C>T~0/1
15.v2_9.v1      42128173~CCTT>C~0/1;42130655~A>AA~0/1;42130715~C>T~0/1   42127307~C>T;42127313~T>C;42128
17.v1_17.v1     42126611~C>G~1/1;42127941~G>A~1/1;42129770~G>A~1/1       42126611~C>G;42127001~G>A;42127
17.v1_19.v1     42126611~C>G~1/1;42127941~G>A~1/1;42128248~CAGTT>C~0/1;42129770~G>A~0/1 42126611~C>G;42
17.v1_22.v1     42126611~C>G~0/1;42127941~G>A~0/1;42129770~G>A~0/1;42130710~G>A~0/1      42126611~C>G;42
17.v1_23.v1     42126611~C>G~0/1;42127941~G>A~0/1;42129770~G>A~0/1;42129836~G>A~0/1      42126611~C>G;42
17.v1_24.v1     42126611~C>G~0/1;42127938~T>G~0/1;42127941~G>A~0/1;42129770~G>A~0/1      42126611~C>G;42
17.v1_25.v1     42126611~C>G~0/1;42127593~G>C~0/1;42127941~G>A~0/1;42129770~G>A~0/1      42126611~C>G;42
17.v1_26.v1     42126611~C>G~0/1;42127514~A>G~0/1;42127941~G>A~0/1;42129770~G>A~0/1      42126611~C>G;42
17.v1_27.v1     42126611~C>G~0/1;42126938~C>T~0/1;42127941~G>A~0/1;42129770~G>A~0/1      42126611~C>G;42
17.v1_28.v1     42126611~C>G~1/1;42127941~G>A~1/1;42129087~G>C~0/1;42129770~G>A~0/1;42130773~C>T~0/1
17.v1_29.v1     42126611~C>G~1/1;42127608~C>T~0/1;42127941~G>A~1/1;42129132~C>T~0/1;42129770~G>A~0/1
17.v1_2.v1      42126611~C>G~1/1;42127941~G>A~1/1;42129770~G>A~0/1       42126310~C>T;42126611~C>G;42126
17.v1_31.v1     42126611~C>G~1/1;42126749~C>T~0/1;42127941~G>A~1/1;42129770~G>A~0/1      42126310~C>T;42
17.v1_32.v1     42126611~C>G~1/1;42126938~C>T~0/1;42127803~C>T~0/1;42127941~G>A~1/1;42129770~G>A~0/1
17.v1_33.v1     42126611~C>G~0/1;42127941~G>A~0/1;42128308~C>A~0/1;42129770~G>A~0/1      42126611~C>G;42
17.v1_34.v1     42126611~C>G~0/1;42127941~G>A~1/1;42129770~G>A~0/1       42126611~C>G;42127001~G>A;42127
17.v1_35.v1     42126611~C>G~1/1;42127941~G>A~1/1;42129770~G>A~0/1;42130761~C>T~0/1      42126310~C>T;42
17.v1_37.v1     42126611~C>G~1/1;42127941~G>A~0/1;42128848~C>T~0/1;42129770~G>A~0/1;42130692~G>A~0/1
17.v1_39.v1     42126611~C>G~1/1;42127941~G>A~0/1;42129770~G>A~0/1       42126611~C>G;42127001~G>A;42127
17.v1_3.v1      42126611~C>G~0/1;42127941~G>A~0/1;42128241~CT>C~0/1;42129770~G>A~0/1     42126611~C>G;42
17.v1_3.v2      42126611~C>G~0/1;42127941~G>A~0/1;42128241~CT>C~0/1;42129042~T>C~0/1;42129770~G>A~0/1
17.v1_41.v1     42126611~C>G~1/1;42127803~C>T~0/1;42127941~G>A~1/1;42129770~G>A~0/1      42126310~C>T;42
17.v1_43.v1     42126611~C>G~0/1;42127941~G>A~0/1;42129770~G>A~0/1;42130715~C>T~0/1      42126611~C>G;42
17.v1_44.v1     42126611~C>G~0/1;42127841~C>G~0/1;42127941~G>A~0/1;42129770~G>A~0/1;42130710~G>A~0/1
:
```

Step 4: Variant combinations corresponding to various diplotypes

Step 5: In parallel, StellarPGx checks for presence of *CYP2D6* copy number variations (CNVs) based on GraphTyper's CNV genotypes as well as read coverage computed using SAMtools bedcov. If no CNV or hybrid rearrangement is detected, StellarPGx assigns the candidate alleles as the final diplotype. For the test case of SIM001, the final diplotype would be *CYP2D6*17/*29*. Some diplotypes, however, have identical combinations of allele defining variants. Classic examples are *CYP2D6*43/*45* and **1/*46*. For such diplotypes, we query different suballele combinations using all the variants in the VCF file. This enables StellarPGx to determine whether a sample's *CYP2D6* variant calls are best explained for example by **43.001/*45.002* instead of **1.006/*46.001.*

```
43.v1_45.v1     43.001_45.001   42126611~C>G~0/1;42127941~G>A~0/1;42129075~C>T~0/1;42130715~C>T~0/1     421
26611~C>G~0/1;42127001~G>A~0/1;42127207~C>T~0/1;42127407~T>G~0/1;42127537~A>G~0/1;42127941~G>A~0/1;42128130
~C>T~0/1;42128216~G>T~0/1;42128662~T>G~0/1;42129075~C>T~0/1;42129130~C>G~0/1;42129950~A>C~0/1;42130047~G>C~
0/1;42130482~C>A~0/1;42130715~C>T~0/1;42131802~A>G~0/1;42131895~C>CT~0/1;42132391~T>A~0/1;42132392~C>A~0/1
43.v1_45.v1     43.001_45.002   42126611~C>G~0/1;42127941~G>A~0/1;42129075~C>T~0/1;42130715~C>T~0/1     421
26611~C>G~0/1;42127001~G>A~0/1;42127207~C>T~0/1;42127407~T>G~0/1;42127537~A>G~0/1;42127941~G>A~0/1;42128130
~C>T~0/1;42128216~G>T~0/1;42129075~C>T~0/1;42129130~C>G~0/1;42129950~A>C~0/1;42130047~G>C~0/1;42130482~C>A~
0/1;42130715~C>T~0/1;42131482~CCACA>C~0/1;42131531~G>A~0/1;42131895~C>CT~0/1;42132026~T>C~0/1;42132089~C>T~
0/1;42132334~C>T~0/1
43.v1_45.v1     43.002_45.002   42126611~C>G~0/1;42127941~G>A~0/1;42129075~C>T~0/1;42130715~C>T~0/1     421
26611~C>G~0/1;42127001~G>A~0/1;42127207~C>T~0/1;42127407~T>G~0/1;42127537~A>G~0/1;42127941~G>A~0/1;42128130
~C>T~0/1;42128216~G>T~0/1;42129075~C>T~0/1;42129130~C>G~0/1;42129950~A>C~0/1;42130047~G>C~0/1;42130482~C>A~
1/1;42130715~C>T~0/1;42131482~CCACA>C~0/1;42131531~G>A~0/1;42131895~C>CT~0/1;42132026~T>C~0/1;42132089~C>T~
0/1;42132334~C>T~0/1
1.v1_46.v1      1.001_46.001    42126611~C>G~0/1;42127941~G>A~0/1;42129075~C>T~0/1;42130715~C>T~0/1     421
26611~C>G~0/1;42127001~G>A~0/1;42127207~C>T~0/1;42127300~C>T~0/1;42127407~T>G~0/1;42127537~A>G~0/1;42127761
~C>T~0/1;42127941~G>A~0/1;42128130~C>T~0/1;42128216~G>T~0/1;42129075~C>T~0/1;42129130~C>G~0/1;42129950~A>C~
0/1;42130047~G>C~0/1;42130482~C>A~0/1;42130715~C>T~0/1;42131531~G>A~0/1;42132026~T>C~0/1;42132089~C>T~0/1;4
2132334~C>T~0/1
1.v1_46.v1      1.001_46.003    42126611~C>G~0/1;42127941~G>A~0/1;42129075~C>T~0/1;42130715~C>T~0/1     421
26611~C>G~0/1;42127001~G>A~0/1;42127207~C>T~0/1;42127407~T>G~0/1;42127537~A>G~0/1;42127941~G>A~0/1;42128130
~C>T~0/1;42128216~G>T~0/1;42129075~C>T~0/1;42129130~C>G~0/1;42129950~A>C~0/1;42130047~G>C~0/1;42130482~C>A~
0/1;42130715~C>T~0/1;42131531~G>A~0/1;42132089~C>T~0/1;42132334~C>T~0/1
1.v1_46.v1      1.006_46.001    42126611~C>G~0/1;42127941~G>A~0/1;42129075~C>T~0/1;42130715~C>T~0/1     421
26611~C>G~0/1;42127001~G>A~0/1;42127207~C>T~0/1;42127300~C>T~0/1;42127407~T>G~0/1;42127537~A>G~0/1;42127761
~C>T~0/1;42127941~G>A~0/1;42128130~C>T~0/1;42128216~G>T~0/1;42129075~C>T~0/1;42129130~C>G~0/1;42129726~A>C~
0/1;42129950~A>C~1/1;42130047~G>C~0/1;42130482~C>A~1/1;42130715~C>T~0/1;42131531~G>A~0/1;42132026~T>C~0/1;4
:
```

Step 5: Example of suballele variant combinations used as tie-breakers when assigning sample diplotypes

However, some diplotypes (e.g. *CYP2D6*102.001/*3.002* and **103.001/*3.001*) cannot be distinguished by StellarPGx's current combinatorial diplotype assignment as they have the exact same variant combination even at suballele level. For such cases, StellarPGx assigns an ambiguous call i.e. *CYP2D6*102/*3 or *103/*3* in this instance.

```
102.v1_3.v2     102.001_3.002   42126611~C>G~0/1;42127941~G>A~0/1;42128241~CT>C~0/1;42129042~T>C~0/1;42129821~
G>A~0/1   42126310~C>T~0/1;42126611~C>G~0/1;42127001~G>A~0/1;42127407~T>G~0/1;42127941~G>A~0/1;42128241~CT>C~0
/1;42129042~T>C~0/1;42129130~C>G~0/1;42129821~G>A~0/1;42130482~C>A~0/1;42130547~T>C~0/1;42130559~T>G~0/1;42130
560~C>G~0/1;42130565~A>G~0/1;42130569~G>C~0/1;42130571~G>T~0/1;42130578~C>G~0/1
103.v1_3.v1     103.001_3.001   42126611~C>G~0/1;42127941~G>A~0/1;42128241~CT>C~0/1;42129042~T>C~0/1;42129821~
G>A~0/1   42126310~C>T~0/1;42126611~C>G~0/1;42127001~G>A~0/1;42127407~T>G~0/1;42127941~G>A~0/1;42128241~CT>C~0
/1;42129042~T>C~0/1;42129130~C>G~0/1;42129821~G>A~0/1;42130482~C>A~0/1;42130547~T>C~0/1;42130559~T>G~0/1;42130
560~C>G~0/1;42130565~A>G~0/1;42130569~G>C~0/1;42130571~G>T~0/1;42130578~C>G~0/1
```

Example with exactly similar variant combinations between two diplotypes, thus leading to ambiguous diplotype assignment.

Supplemental note:

Evaluating StellarPGx's performance on 35,778 diplotype simulations

In addition to our benchmark using publicly available WGS data, we tested the ability of StellarPGx's combinatorial diplotype assignment to resolve 35,778 diplotype simulations (all combinations of 267 SNV-defined haplotypes covering *CYP2D6*1.001 – *140.001* from PharmVar v4.2.4 (`https://www.pharmvar.org/gene/CYP2D6`) corresponding to GRCh37). All the haplotypes used in the simulations are listed in Table S4.2.

Firstly we generated all diplotype combinations of the test haplotypes in VCF format using an in-house Nextflow script, get_dips.nf and a custom Python 3.8 script, sim_zygosity.py (see StellarPGx repository `https://github.com/SBIMB/StellarPGx/tree/master/tests/simulations`).

We then filtered out duplicate diplotypes and ran each VCF file through our combinatorial *CYP2D6* diplotype assignment script (see `https://github.com/SBIMB/StellarPGx`).

Of the 35778 test cases, 35764 diplotypes were correctly called (Table S4.2). However, we had 14 cases where StellarPGx reported ambiguous calls (Table S4.2). These ambiguous calls were due to cases where two diplotypes had the same variant combinations (including core and synonymous SNVs). The classic example of *CYP2D6*102.001/*3.002* and **103.001/*3.001* has already been presented above.

References

1. Gaedigk, A., *et al*. (2018). The Pharmacogene Variation (PharmVar) Consortium: incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clinical Pharmacology and Therapeutics*. 103 (3), 399–401.

2. Robinson, J. T. *et al*. (2011). Integrative genomics viewer. *Nature Biotechnology*. 29 (1), 24–26.

3. Eggertsson, H. *et al*. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*. 49, 1654–1660.

Appendix E: Published papers

npj | Genomic Medicine

www.nature.com/npjgenmed

ARTICLE OPEN

A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on *CYP2D6* genotyping

David Twesigomwe[1,2]✉, Galen E. B. Wright[3,4], Britt I. Drögemöller[5], Jorge da Rocha[1,2], Zané Lombard[2] and Scott Hazelhurst[1,6]✉

Genetic variation in genes encoding cytochrome *P450* enzymes has important clinical implications for drug metabolism. Bioinformatics algorithms for genotyping these highly polymorphic genes using high-throughput sequence data and automating phenotype prediction have recently been developed. The *CYP2D6* gene is often used as a model during the validation of these algorithms due to its clinical importance, high polymorphism, and structural variations. However, the validation process is often limited to common star alleles due to scarcity of reference datasets. In addition, there has been no comprehensive benchmark of these algorithms to date. We performed a systematic comparison of three star allele calling algorithms using 4618 simulations as well as 75 whole-genome sequence samples from the GeT-RM project. Overall, we found that Aldy and Astrolabe are better suited to call both common and rare diplotypes compared to Stargazer, which is affected by population structure. Aldy was the best performing algorithm in calling *CYP2D6* structural variants followed by Stargazer, whereas Astrolabe had limitations especially in calling hybrid rearrangements. We found that ensemble genotyping, characterised by taking a consensus of genotypes called by all three algorithms, has higher haplotype concordance but it is prone to ambiguities whenever complete discrepancies between the tools arise. Further, we evaluated the effects of sequencing coverage and indel misalignment on genotyping accuracy. Our account of the strengths and limitations of these algorithms is extremely important to clinicians and researchers in the pharmacogenomics and precision medicine communities looking to haplotype *CYP2D6* and other pharmacogenes using high-throughput sequencing data.

npj Genomic Medicine (2020)5:30 ; https://doi.org/10.1038/s41525-020-0135-2

INTRODUCTION

Genetic variation is known to influence the way in which individuals respond to therapeutics. Groups of variants that are inherited together, known as haplotypes, provide a basis for phenotype prediction and treatment decisions during pharmacogenomic testing. Accurately detecting functional haplotypes, popularly known as star (*) alleles, in clinically actionable pharmacogenes (e.g. cytochrome *P450* genes) is therefore crucial to the implementation of personalised medicine. In addition, determining pharmacogene star allele frequencies and consequently predicting the phenotypic landscape is a crucial part of large-scale population genetic studies, which can help inform drug policy[1].

Of the clinically relevant pharmacogenes, *CYP2D6* is one of the most widely studied owing to its contribution to the Phase 1 metabolism of ~25% of clinically prescribed drugs, and high genetic variability within and between populations[2]. Drugs metabolised by *CYP2D6* include various antidepressants, antipsychotics, anticancer agents (e.g. tamoxifen), and opioids (https://drug-interactions.medicine.iu.edu/MainTable.aspx, last accessed: 19-02-2020).

The highly polymorphic *CYP2D6* gene has nine exons and is located on chromosome 22q13.2 neighbouring two paralogous pseudogenes, *CYP2D7* and *CYP2D8*, which further complicates genotyping[3,4]. To date, the Pharmacogene Variation (PharmVar) Consortium[5] has catalogued over 130 different *CYP2D6* star alleles with varying levels of evidence (https://www.pharmvar.org, last accessed: 19-02-2020). The majority of *CYP2D6* star alleles are defined by specific combinations of single nucleotide polymorphisms (SNPs) and/or small insertions and deletions (indels) (collectively referred to as single/small nucleotide variants (SNVs) for the rest of this manuscript), which may either alter the function of the protein or be neutral. In addition, the *CYP2D6* gene locus contains a number of complex structural variants including full gene deletions, gene duplications and multiplications, as well as hybrid tandem rearrangements with the highly similar *CYP2D7*[6–8]. These allelic variants are graphically shown in Fig. 1.

The complexities in the *CYP2D* gene locus present a significant challenge in accurately genotyping *CYP2D6*. Currently, gold standard genotyping involves a battery of methods e.g. quantitative PCR analysis for CNV detection, allele specific XL-PCR, SMRT sequencing, and Sanger sequencing. However, these techniques are expensive, laborious and unscalable, especially for large cohort studies[9–11]. Further, although more economical, selectively targeting a subset of alleles through individually genotyping defining variants can have important clinical implications as inaccurate phenotype prediction is more likely to occur[11,12]. The widespread availability of next-generation sequencing (NGS) characterised by high-throughput parallel DNA sequence data generation provides a solution to the cost-limitations of experimental methods for genotyping *CYP2D6* and other pharmacogenes. The ever-decreasing cost of NGS facilitates whole-genome

[1]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa. [2]Division of Human Genetics, National Health Laboratory Service, and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [3]Neuroscience Research Program, Kleysen Institute for Advanced Medicine, Winnipeg Health Sciences Centre and Max Rady College of Medicine, University of Manitoba, Winnipeg, MB, Canada. [4]Department of Pharmacology and Therapeutics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada. [5]Department of Biochemistry and Medical Genetics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada. [6]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa. ✉email: twesigomwedavid@gmail.com; scott.hazelhurst@wits.ac.za

Published in partnership with CEGMR, King Abdulaziz University

npj nature partner journals

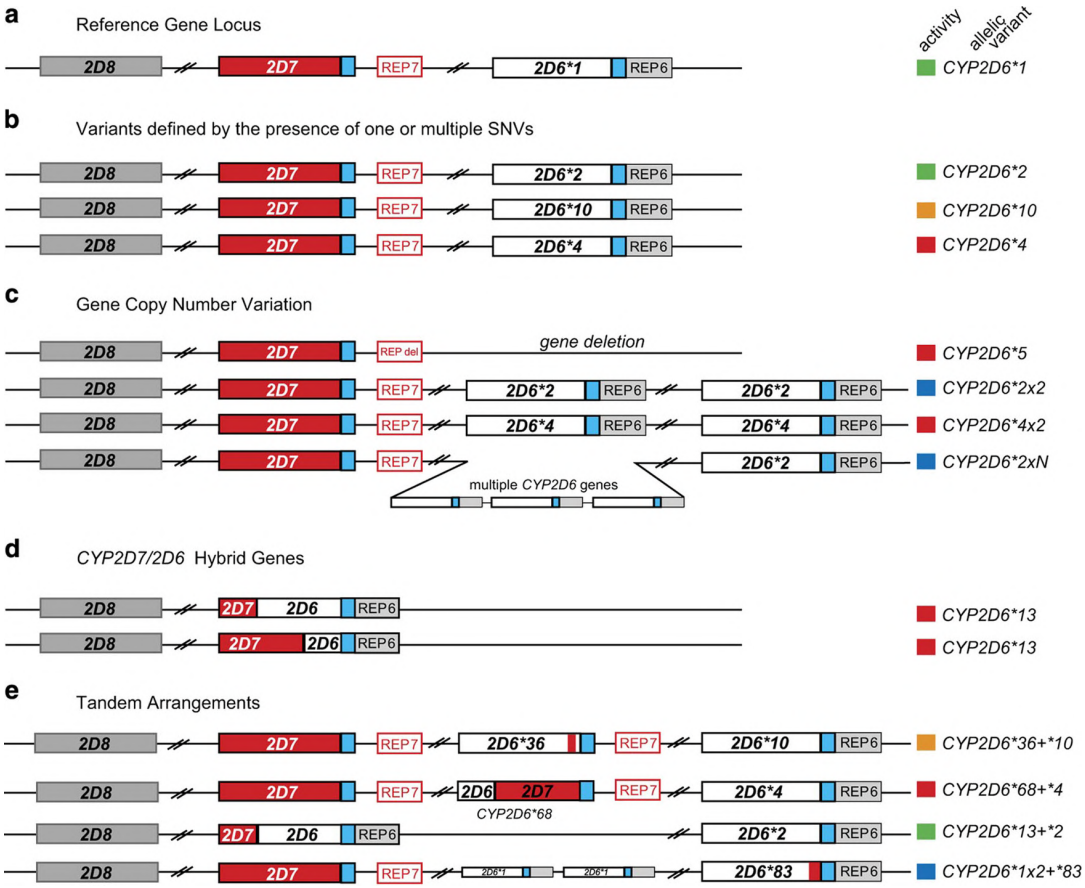


Fig. 1 Graphical overview of the highly polymorphic *CYP2D6/2D7/2D8* locus, by Twist et al.[15], licensed under Creative Commons CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). No changes have been made to the figure content. a The relative position of the reference *CYP2D6*1* haplotype (white) to two non-functional paralogs, *CYP2D7* (red) and *CYP2D8* (grey) on the minus strand of Chromosome 22. REP6 and REP7 are paralogous, Alu-containing, 600-bp repetitive sequences found downstream of *CYP2D6* and *CYP2D7*, respectively. The blue boxes indicate identical unique sequences downstream of *CYP2D6* and *CYP2D7*. Notice the "spacer" (1.6-kb) separating REP7 from *CYP2D7* but none between *CYP2D6* and REP6. **b** Common *CYP2D6* star alleles defined by core single nucleotide variants (SNVs). **c** Examples of *CYP2D6* copy number variations and their functional annotation. **d** Examples of *CYP2D7/2D6* hybrid genes. **e** Common tandem rearrangements in the *CYP2D* gene locus. The activity level boxes on the right are coded; red for a non-functional haplotype, orange for decreased activity, green for fully functional reference activity, and blue for increased activity.

sequencing (WGS) in addition to providing options for the development of gene panels such as PGRNseq[13].

In addition, NGS methods have the advantage of being unbiased, thus allowing for the detection of rare and novel variations, which are more likely to be deleterious[14]. Although there are important benefits associated with NGS-based genotyping, these approaches present computational challenges, mainly relating to the alignment of short-read data in regions of high sequence similarity. This makes variant calling and haplotyping genes such as *CYP2D6*, that are located in regions of high sequence similarity, challenging.

To address these challenges, bioinformatics tools that have recently been developed for calling star alleles in *CYP2D6* and other highly polymorphic pharmacogenes using genome sequencing data and/or targeted-capture panels such as PGRNseq, include Astrolabe (formerly Constellation)[15], Aldy[16], Stargazer[17], VCF Annotator[18], Cypiripi[19], and PharmCAT[20]. These tools automate the detection of diplotype combinations based on PharmVar and the Pharmacogenomics Knowledgebase (PharmGKB) star allele catalogues thus facilitating clinical interpretation. The importance of these algorithms in *CYP2D6* phenotype prediction in clinical settings, and allele discovery in research studies, therefore cannot be overstated. However, to date, there is no comprehensive comparison of these tools. Astrolabe, Aldy, Stargazer, and PharmCAT are regularly maintained. However, PharmCAT does not perform star allele calling for *CYP2D6* directly, but rather uses Astrolabeś allele calling output for its unique clinical annotation step, which is based on current clinical implementation guidelines[20,21].

We therefore provide a much-needed in-depth comparison of the performance of Astrolabe, Aldy, and Stargazer on a wide range of *CYP2D6* allelic variation using simulated data (entire *CYP2D* locus) and real data. The research participant-derived data used in the study are part of the Polaris pharmacogenomics cohort (https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort#Pratt2016) whose samples were originally collected as part of the HapMap and 1000 genomes projects[22]. The Centers for Disease Control and Prevention (CDC)-based Genetic Testing Reference Material Coordination Program (GeT-RM) has characterised these samples through extensive orthogonal testing by various laboratories[23,24]. In doing so, GeT-RM has contributed invaluable reference materials for benchmarking *CYP2D6* genotyping approaches.

In addition, we evaluate the impact of read depth on recall of *CYP2D6* CNVs for each tool. Given the differences in the allele calling approach of these algorithms, we further analyse the

Table 1. Properties of Astrolabe, Aldy, and Stargazer.

Tool	OS	Language	Central feature	NGS data	Input/ref	Output
Astrolabe	Linux Mac	Java	Probabilistic scoring system	WGS PGRNseq[a]	VCF BAM b37, b38	Diplotypes Suballeles Phenotype All novel SNVs
Aldy	Linux Mac	Python	Combinatorial framework	WGS PGRNseq	BAM b37	Diplotypes Suballeles Putative novel core variants
Stargazer	Linux Mac	Python	Statistical phasing	WGS PGRNseq	VCF GDF b37	Diplotypes Phenotype All novel SNVs

[a]Not yet optimised for Astrolabe's CNV calling.

efficacy of an ensemble genotyping approach involving all three tools using the real data.

Through this comparative analysis, we provide important information on the strengths and limitations of current pharmacogene star allele calling algorithms. Our recommendations will serve as a valuable resource for the pharmacogenomics community regarding the use of these algorithms in clinical and research settings, and provide important data for developers aiming to improve the software or develop new ones. A summary of the main features of Astrolabe, Aldy, and Stargazer is given in Table 1.

RESULTS

Criteria

We evaluated Astrolabe, Aldy, and Stargazer using three criteria including;

- *Haplotype concordance*: The percentage of total haplotype calls that are consistent with the ground truth. The haplotype true positive rate is also known as the analytic specificity.
- *Diplotype (genotype) concordance*: The percentage of diplotype calls that are consistent with the ground truth (i.e. correctly calling both haplotypes for each sample).
- *Phenotype concordance*: Percentage of assigned phenotypes (based on activity scores) concordant to the reference materials.

The results reported here are specific to Astrolabe (v0.8.7.0), Aldy (v2.2.3), Stargazer (v1.0.7), and the datasets (simulated and real) used in this study. The goal was to have datasets with a wide variety of catalogued *CYP2D6* alleles. The simulated data (4618 test cases) do not represent actual population frequencies. In comparison, the GeT-RM WGS data (75 samples) are representative of African, Admixed American, European, and Asian allele distributions.

Simulated datasets

Set 1 of our simulations comprised all theoretically possible diplotype combinations (homozygous and heterozygous) derived from PharmVar-catalogued *CYP2D6* SNV-defined haplotypes. Set 2 comprised diplotypes with at least one structural variant (*CYP2D6* gene deletion (**5*), duplications, exon conversions and hybrid rearrangements i.e. *CYP2D6/2D7* and *CYP2D7/2D6* hybrids). These structural variations are comprehensively described by PharmVar (https://www.pharmvar.org/gene-support/Variation_CYP2D6.pdf). For a complete list of all the *CYP2D6* haplotypes used for generating simulations in this study, see Supplementary Data Set 1.

Performance based on simulated data

In set 1 (SNV-defined alleles), the concordance of Astrolabe and Aldy for homozygous diplotypes was comparable as they correctly identified 142 (92%) and 148 (96%) diplotypes, respectively, out of 154 diplotypes (Table 2, Run 1). However, Stargazer had lower concordance (135 i.e. 88%) for the homozygous diplotypes. On further investigation, we found 18 cases where Stargazer prioritised reporting background functionally annotated alleles as the main result for haplotypes that have uncertain function. For example, Stargazer reported *58/*58 as *17/*17 given that *58 (unknown function) and *17(decreased function) both have the T107I variant. In such cases, Stargazer reported the expected allele (s) ambiguously with other candidate solutions (see Table 2, Supplementary Data Set 2, and Supplementary Information for additional details).

Calling some indel-defined alleles was challenging for all the tools due to left/right shifts in the alignments. These shifts emphasise challenges of using short-read NGS data for genotyping *CYP2D6* as they caused position discordance in the input data (BAM and/or VCF) and the algorithms' definitions (see Table 2 for examples of problematic indel-defined alleles). Discordant cases for Astrolabe and Aldy involved undefined suballeles as well (Table 2), resulting in ambiguous genotypes or miscalls. The concordance of Astrolabe and Aldy increased to 99% and 97%, respectively, after updating their allele databases with previously undefined suballeles (Table 2, Run 2).

We next considered tool concordance on all theoretically possible diplotype combinations of *CYP2D6* SNV-defined haplotypes with definitive or moderate PharmVar level of evidence ($N = 4560$ i.e. 95 choose 2 plus 95). The main diplotypes called by all three tools are shown in Supplementary Data Set 3. The concordance of Astrolabe and Aldy was again comparable (99% and 97%, respectively), whereas Stargazer was concordant for only (43%) of the possible diplotypes (Fig. 2). Stargazer's lower recall can be attributed to not only the aforementioned differences in reporting uncertain function alleles but also to the limitations of the statistical phasing process central to the Stargazer pipeline. For example, the *74/*48 combination involving a rare African-specific haplotype (*74) and a rare East Asian-specific haplotypes (*48) was called as *1/*1 by Stargazer. Homozygous cases of these alleles are correctly called by Stargazer via manual phasing. However, rare heterozygous combinations are more difficult to call since population structure affects statistical phasing. Notably, some alleles are correctly called using

Table 2. Concordance of Astrolabe, Aldy, and Stargazer on test cases ($N = 154$) homozygous for SNV-defined star alleles at 30×.

	Astrolabe		Aldy		Stargazer
	Run 1	Run 2	Run 1	Run 2	
Match	142 (92%)	152 (99%)	148 (96%)	149 (97%)	135 (88%)
Mismatch	9 (6%)	1 (1%)	4 (3%)	3 (2%)	19 (12%)
Ungenotyped	3 (2%)	1 (1%)	2 (1%)	1 (1%)	0

Inconsistencies

Algorithm	Challenges	Discordant cases	Notes
Astrolabe	Undefined suballeles	*2.014, *4.002, *4.003, *4.007, *4.009, *52.002, *56.003	Ambiguous calls or miscalls
	Indel-defined alleles	*18.001	Position discordance with database
Aldy	Undefined suballeles	*12.002, *71.003	Ambiguous calls or miscalls
	Indel-defined alleles	*18.001, *20.001, *38.001	Position discordance with database
Stargazer	Indel-defined alleles	*20.001	Position discordance with database
	Alleles with unknown or uncertain function and/or rare alleles	18 alleles including *30.001, *37.001, *58.001, *52.001, *64.001, *65.001, and *73.001 (Supplementary Data Set 2)	Stargazer reports background alleles as the main haplotypes in these cases. Exact haplotypes (matching truthset) in the sample(s) are reported among other candidate haplotypes. (See discussion for details)

The complete list of genotypes called per sample is provided in Supplementary Data Set 2.
Run 1 and Run 2 indicate results before and after defining missing suballeles in the allele tables, respectively. There was only one run for Stargazer as it does not typically call suballeles.

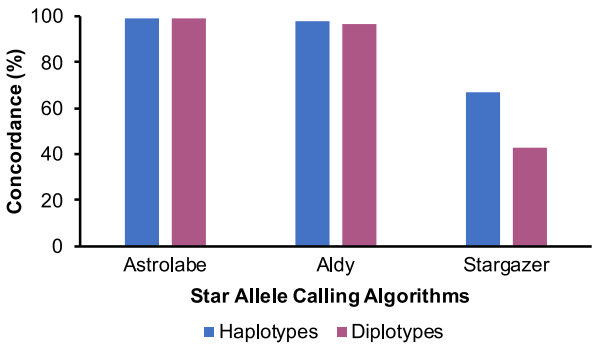


Fig. 2 Concordance of Astrolabe, Aldy, and Stargazer for all theoretically possible *CYP2D6* diplotype combinations (homozygous and heterozygous) comprising SNV-defined haplotypes with PharmVar definitive or moderate level of evidence from our starting set. We left the allele databases of the three tools as is in order to examine the effect of the undefined suballeles on diplotype calling. The sequencing coverage for each test case was 30× and default parameters were used for each tool.

Stargazer's "phasing by haplotype extension algorithm" that complements the statistical phasing (https://stargazer.gs.washington.edu/stargazerweb/). Astrolabe and Aldy do not do statistical phasing; hence, they are virtually not affected by population structure.

For all three tools, the concordance for haplotypes was much higher than that for diplotypes. This underscores the challenge of consistently calling both haplotypes for each sample in a highly polymorphic gene such as *CYP2D6*. We show the true extent of this later on with the real data that follows true population distributions.

All three tools accurately genotyped simulated samples with at least one *CYP2D6* full gene deletion allele (*5) (i.e. CN = 0,1) as indicated in Table 3. The notable discordances were samples that contained a duplicated allele on the second haplotype (e.g. *5/*29×2 was called as *29/*29). Astrolabe in particular also had challenges calling the *5/*5 diplotype (CN = 0). Astrolabe reported failed BAM quality control for the *5/*5 test cases. Running the algorithm by skipping quality control unsurprisingly yielded haplotype calls matching GRCh37 (i.e. *2M/*2M, indicating absence of non-reference SNVs). Varying the coverage from 30× to 100× had very little effect on the *CYP2D6*5* recall of Aldy and Stargazer. However, Astrolabe detected 5 from 20 *CYP2D6*5* alleles at 60× compared to 10 from 20 at 30× and 100×. The miscalls were observed mainly in samples with CN = 1.

For copy number gains (CN > 2), Aldy and Stargazer outperformed Astrolabe by detecting up to 50% more duplication/multiplication events. By default, Aldy and Stargazer resolve copy number gains further to report which of the star alleles was duplicated or multiplicated. However, for our simulated test cases with high copy number, some discordant tandem arrangements were reported by Aldy and Stargazer. For example, Aldy genotyped the test case *1×4/*2×8 as *1×3+*63/*2×7+*79, whereas Stargazer genotyped it as *1×6/*34×6. Genotypes with such high *CYP2D6* copy number are rare but possible as shown previously[25]. It is worth mentioning that Aldy was the only algorithm that resolved the *90+*1 non-hybrid tandem arrangements.

In contrast to Stargazer and Aldy, Astrolabe does not actively resolve duplicated/multiplicated alleles but rather generically indicates that a gene duplication has been detected. Similarly by default, Astrolabe does not distinguish between a duplication and a multiplication but rather generically indicates the presence of a gene duplication if copy number gain is detected. The concordance for duplications/multiplications was more or less equally good at 30×, 60× and 100× for Aldy and Stargazer (Table 3). For Astrolabe, there was no evidence to suggest that increasing coverage from 30× to 60× or 100× increases recall of duplications. This could be due to limitations of using short-read NGS data for interrogating the complex *CYP2D6* region.

Regarding gene exon conversions and hybrid rearrangements, Aldy consistently called more of such alleles at 30× (72 out of 107) compared to Stargazer (31 out of 107) and Astrolabe (3 out of

Table 3. Summary of correctly called structural variations by Astrolabe, Aldy, and Stargazer in set 2.

Set 2	Truth	Astrolabe			Aldy			Stargazer		
		30×	60×	100×	30×	60×	100×	30×	60×	100×
Full gene deletion (*5)	20	10	5	10	15	14	15	15	15	15
Copy number gain[a]	52	29	25	25	52	50	52	47	45	45
Resolved duplicated/ multiplicated alleles	61	4[c]	4[c]	4[c]	53	49	51	54	51	48
Hybrids[b]	107	3	8	8	72	66	67	31	30	28
Non-hybrid tandem	4	0	0	0	4	4	4	0	0	0
Ungenotyped (defaults)		0	0	0	1	3	2	18	16	19

The complete list of genotypes called per sample is provided in Supplementary Data Set 4.
[a]Due to gene duplications/multiplications.
[b]Collectively representing exon conversions and gene hybrids involving *CYP2D6* and *CYP2D7*.
[c]Determined only for homozygous allele cases.

107). Astrolabe consistently called *82 at 30×, 60×, and 100× unlike Aldy and Stargazer, and then consistently called the *68+*4 tandem only at 60× and 100×. As observed for duplication events, increase in coverage from 30× to 60× and 100× did not considerably affect Aldy and Stargazer's performance for calling hybrids (see Table 3 and Supplementary Data Set 4 for additional details).

Comparison of *CYP2D6* star allele calling performance on real data

To evaluate the genotyping accuracy of Aldy, Astrolabe, and Stargazer on real data, we applied all three algorithms to 75 ethnically diverse samples from the GeT-RM project. GeT-RM previously characterised 137 DNA samples from Coriell cell lines for 28 genes, including *CYP2D6*, through targeted genotyping by selected laboratories[23]. GeT-RM has recently re-characterised these 137 samples and genotyped 42 new samples to ascertain complex rearrangements and rare variants in *CYP2D6*[24]. Of the combined 179 samples, 75 currently have publicly available high coverage WGS data (see Data availability section for details). We used published consensus *CYP2D6* genotypes[23,24] for these samples as the Gold Standard calls except for sample NA18519. GeT-RM used only panel-based genotyping for this sample resulting in the *1/*29 diplotype call. However, the *106 defining variant 3878G>A as well as the *29 defining variants were detected from NGS. Therefore, *106/*29 is considered to be the true positive call for sample NA18519 for this analysis.

When applied to the GeT-RM WGS samples, both Astrolabe and Stargazer assigned diplotypes for 75 of 75 samples. One sample could not be genotyped by Aldy with default parameters. Stargazer had the highest genotype concordance (89%—67 of 75 diplotypes) to the ground truth followed by Aldy (88%—66 of 75 diplotypes) and then Astrolabe (72%—54 of 75 diplotypes). The haplotype concordance followed a similar trend i.e. 94% (141 of 150 haplotypes), 91% (136 of 150 haplotypes), and 83% (125 of 150 haplotypes) for Stargazer, Aldy, and Astrolabe, respectively (Fig. 3a).

Notably, the ensemble genotyping approach (2 from 3 rule) had the highest diplotype concordance (95%—71 of 75 diplotypes) and haplotype concordance (97%—145 of 150 haplotypes).

Samples with *CYP2D6* CNVs were more challenging to call for all the algorithms notably Astrolabe (15 of 75 samples) (Fig. 3b). Astrolabe particularly had challenges calling *36, a *CYP2D6*/2D7 hybrid gene that frequently occurs singly or in tandem with *10. Tandem rearrangements involving *36 present in the GeT-RM dataset include *36×2, *36+*10, and *36×2+*10. Aldy and Stargazer consistently detected the more common *36+*10 arrangement but could not call all copies correctly for the more complex *36×2 and *36×2+*10 arrangements (see Supplementary Data Set 5 for detailed genotype calls).

All the three tools consistently called the *68+*4 tandem arrangement except for one case where Astrolabe called sample NA21781 as *2/*4 (possible duplication) instead of *2×2/*68+*4. *68 is a non-functional *CYP2D6*/2D7 hybrid gene with an intron 1 breakpoint. The *68+*4 haplotype is therefore non-functional as *4 is also a non-functional allele.

As shown in Fig. 3c, haplotype calls by one tool that are not confirmed by either of the other two tools are mostly incorrect. The caller overlap between Aldy and Stargazer was more prominent than either tool's intersection with Astrolabe because of Astrolabe's lower recall for structural variations. The ensemble genotyping approach reduced the number of discordant genotype calls especially for samples "without" *CYP2D6* CNVs (many of these have haplotypes with the neutral intron 1 conversion). However, some alleles defined by *CYP2D6* CNVs (especially aforementioned *36 hybrid arrangements) proved to be challenging even with ensemble calling (Fig. 3b). In addition, the ensemble genotyping approach had two ambiguous calls (from samples NA19908 and NA18540) arising from discrepant calls from all three tools (see Supplementary Data Set 5).

Comparison of *CYP2D6* phenotype prediction concordance

We next evaluated the clinical accuracy of Astrolabe, Aldy, Stargazer, and the ensemble approach. We followed the current consensus clinical implementation guidelines[21,26] to assign activity scores corresponding to the genotype calls and compared them to the GeT-RM consensus reference. All the three algorithms had much higher phenotype concordances compared to their diplotype concordances (Fig. 4). This is due to the fact that even though some samples may not be accurately genotyped, the activity score or phenotype group of the reported diplotype could be the same as that of the truth call. For instance, sample NA18565 was called as *10/*36+*10 (Stargazer and Aldy) and *10/*10 (Astrolabe) instead of *10/*36×2 (truth call). All these genotypes denote an intermediate metaboliser phenotype. Aldy (95%) and Stargazer (96%) had comparably higher phenotype concordance than Astrolabe (91%) because they were able to consistently genotype more samples with *CYP2D6* CNVs as mentioned earlier. Notably, the ensemble approach had the highest phenotype concordance (97%) and it was comparable to its diplotype concordance (95%). In some cases, we obtained phenotype "no calls" due to ungenotyped samples e.g. for Aldy(1) and the ensemble approach (two ambiguous cases).

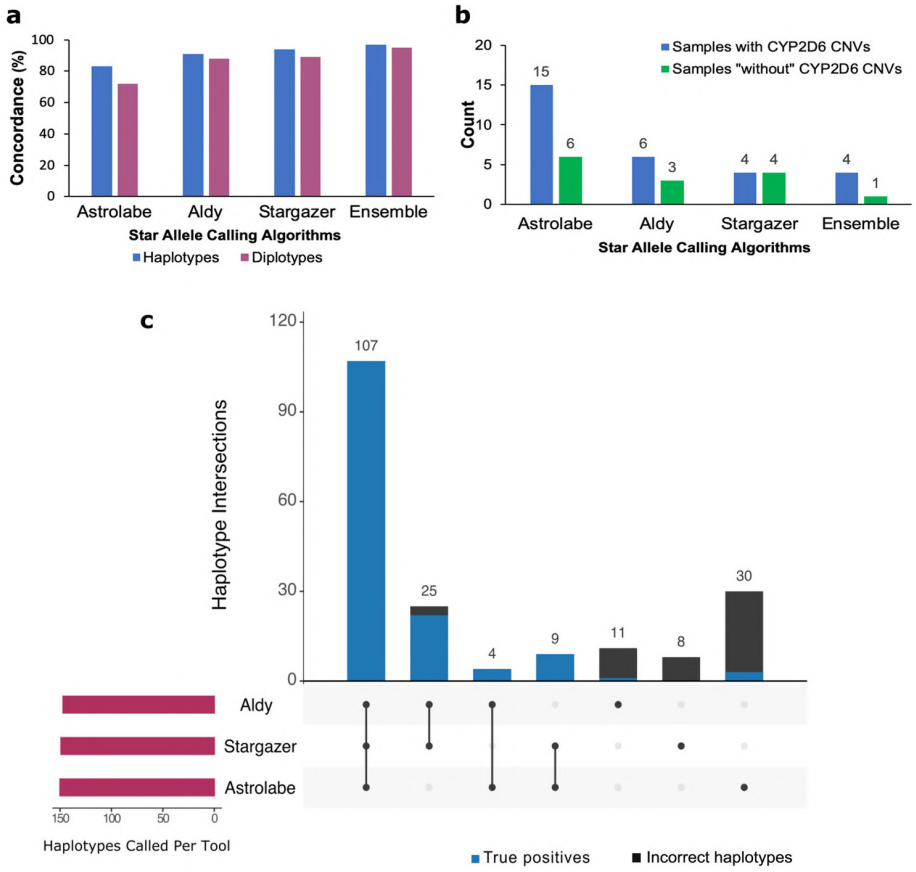


Fig. 3 Performance of Astrolabe, Aldy, and Stargazer on 75 WGS GeT-RM samples. a Concordance of each algorithm to the GeT-RM consensus calls. **b** Cases with discordant genotypes. The green colour represents samples "without" *CYP2D6* CNVs (most of the samples may have the intron 1 conversion), whereas the blue colour represents samples with allele-defining *CYP2D6* CNVs. **c** Overlap of haplotypes called by each algorithm. As shown ensemble call sets have high concordance. Haplotypes called by one tool but not confirmed with any of the others have low true positive rate.

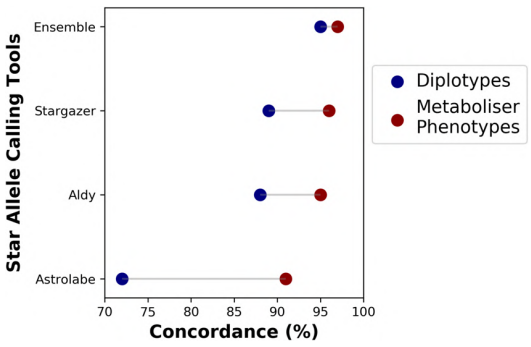


Fig. 4 Comparison between the CYP2D6 diplotype concordance and the phenotype prediction concordance for each algorithm based on 75 GeT-RM samples and the Activity Score system. All three algorithms and the ensemble approach have relatively high phenotype concordance thus underscoring their clinical utility even for some cases with inconsistent diplotype calls.

DISCUSSION

Accurately calling star alleles in *CYP2D6* is crucial to inferring phenotype from genotype. However, *CYP2D6* is one of the most challenging genes to genotype largely because of the complexity of its genomic locus. The limitations of using short-read NGS data compound the problem even further. Herein, we have benchmarked three algorithms that use largely different approaches to call *CYP2D6* star alleles from high-throughput sequence data.

We used simulations to test the performance of each algorithm on a wide variety of *CYP2D6* allelic variation i.e. 154 SNV-defined haplotypes (many have the *CYP2D7* intron 1 conversion), and over 50 different structural variant arrangements (Supplementary Data Set 1). We also assessed the accuracy of each algorithm using real WGS data from CDC-based GeT-RM project.

Overall, Astrolabe and Aldy call more combinations of SNV-defined *CYP2D6* star alleles (homozygous and heterozygous) compared to Stargazer. Stargazer faces a challenge in calling rare (frequency < 0.1%) SNV-defined alleles due to its reliance on statistical phasing. Although Stargazer has supplementary algorithms to phase rare variants that are not present in the reference panel, we found that they work best in homozygous cases but have lower accuracy in heterozygous cases. Stargazer also had discordances due to its difference in reporting alleles with unknown/uncertain function. For such alleles, Stargazer currently prioritises to report background alleles for which the function is known as the main haplotypes. However, it is difficult to infer the exact alleles in the sample as they are reported in a separate column ambiguously with other candidate alleles. Notably, homozygous cases of these alleles are assigned unfounded phenotypes. For example, *58/*58 (unknown metaboliser status)

is called as *17/*17 (intermediate metaboliser) by Stargazer, whereas *58 is reported among candidate alleles for haplotype 1 and 2.

Notably all three algorithms depend on the comprehensive definition of alleles (and suballeles especially for Astrolabe and Aldy) to avoid ambiguous calls and miscalls. Indel-defined star alleles are a major challenge for all three algorithms especially where the core SNV occurred in a repeat region causing the aligner to pick the left-most possible position in the alignment (BAM) file. Core indels for *18*, *20*, *38*, *40*, and *42* all appear in unexpected positions due to this problem. Miscalling these alleles could have potential clinical implications. Even though the three algorithms have corrected for some of these changes in their allele definitions, the complexity of their locus can affect coverage and variant quality scores. Of note, other alleles with indel core variants such as *3, *6, *9, *19, and *21 were consistently called.

Regarding the calling of *CYP2D6* structural variants, duplications/multiplications and deletions were considerably easier to call compared to hybrid rearrangements. Aldy had the highest concordance followed by Stargazer, and then Astrolabe. Notably, Astrolabe does not distinguish between duplications and multiplications (i.e. identifies all scenarios as duplications), and also does not resolve the exact duplicated gene copy (or copies) for samples where it detects copy number gains. This could affect phenotype prediction depending on the diplotype combination in a sample. Astrolabe also had challenges detecting the *5/*5 diplotype (CN = 0) probably because its current model for structural variant calling was not trained on enough cases with zero copy number. In addition, for hybrid rearrangements, Astrolabe consistently detected only *82* (has an exon 2 conversion) and *68+*4* (at 60× and 100×). As with the duplication events, the effect of Astrolabe's inconsistency in calling hybrid rearrangements on phenotype prediction depends on the entire diplotype combination of an individual.

Stargazer had the intermediate performance for calling structural variants of the three tools. It was able to phase CNVs and detect more hybrid rearrangements compared to Astrolabe while also producing visual plots (Supplementary Fig. 1) of the structural variants in the process. However, Stargazer had the highest number of indeterminate (i.e. ungenotyped) diplotypes for samples with structural variants. Notably, Stargazer still produced copy number and allele fraction plots for these samples. However, ungenotyped calls can be a major hindrance to automated phenotype prediction in practice.

Aldy performed best in calling *CYP2D6* structural variants. However, as with Astrolabe and Stargazer, it had challenges in reconstructing complex hybrid rearrangements such as *79+*68+*4* and *79+*68+*4×2*, and it also incorrectly indicated the presence of hybrids such as *63* and *79* for samples with high copy number. Although high copy number diplotypes such as *1×4/*2×8 are rare, it is important to examine whether they can be picked up by these algorithms. This is particularly important given that some populations are largely understudied and may have individuals with such alleles.

To evaluate the performance of Astolabe, Aldy, and Stargazer on real data, we used 75 sequences from the 1000 Genomes Project for Coriell reference samples that have been re-characterised through the GeT-RM project. Astrolabe and Stargazer were able to call diplotypes for 100% of the samples, whereas one sample could not be diplotyped using Aldy's default parameters. However, the concordance of Aldy and Stargazer was comparably higher than that for Astrolabe mainly due to their superiority in calling *CYP2D6* CNVs. Although the 75 GeT-RM WGS samples used for this analysis depicted the complexity and diversity in *CYP2D6* pharmacogenomic variation, they did not include a number of rare alleles. Future availability of WGS data for the remaining 104 GeT-RM reference materials will provide the opportunity to test these algorithms against a more comprehensive distribution of *CYP2D6* star alleles.

As with the simulated data, gene deletions as well as duplications/multiplications were more consistently called compared to hybrid rearrangements. However, Astrolabe missed more deletions and duplications than Aldy and Stargazer did. This could be attributed to Aldy and Stargazer's coverage normalisation strategies during structural variant detection that are not utilised by Astrolabe.

Regarding hybrids, the *68+*4 non-functional tandem arrangement was consistently called by all three algorithms. Notably, Astrolabe could not differentiate *36 from *10. On the other hand, Aldy and Stargazer consistently called the *36+*10 arrangement but could not differentiate more complex arrangements such as *36×2, *36×2+*10. This can be attributed to incomprehensive definition of these haplotypes within the algorithms and also the limitations of using short-read NGS data.

We also assessed the overlap of Astrolabe, Aldy, and Stargazer calls as well as the practical application of ensemble *CYP2D6* star allele calling. Simple ensemble calling could be useful in obtaining high-confidence *CYP2D6* star allele calls both in the clinical setting and in large-scale population studies. Through obtaining high-confidence star allele calls, the number of samples requiring experimental validation (e.g. those found to have putative novel alleles and structural variants) would be reduced to a smaller subset of the study population. We generated the ensemble call set by taking a consensus of Aldy, Astrolabe and Stargazer allele calls for each sample on a 2 from 3 rule. We obtained 100% concordance for haplotypes (107) that were called by all three algorithms, and only 5 discordant genotypes compared to Astrolabe (21), Aldy (9), and Stargazer (8) from 75 samples. Based on the challenges and limitations of each algorithm and of using short-read NGS data, we recommend the use of all three algorithms to perform highly accurate ensemble calling. The ensemble calling is also practical as the typical runtime for each of three algorithms from our workflows was less than 1 min per sample while using 4 CPU cores.

Although it is quite straightforward to obtain high-confidence SNV-defined star allele calls using ensemble calling, it is more challenging to get unambiguous high-confidence diplotype calls for individuals with complex structural variations especially the hybrid rearrangements. This follows from the differences in accuracy of the three algorithms for different forms of *CYP2D6* allelic variation described above. A worry with ensemble calling in the clinic would be how to resolve the discrepancies among the tools. These samples would require orthogonal testing to resolve the number and exact sequence content of their *CYP2D6* copies. Further work is required to build an automated ensemble allele calling and phenotype prediction pipeline. This will heavily depend on achieving uniformity in the precision of reporting star alleles and suballeles by all the algorithms in their future updates.

All three algorithms had much higher phenotype concordance compared to their diplotype concordance implying that their clinical accuracy does not necessarily match their genotyping accuracy. Notably, the ensemble genotyping approach had comparably high phenotype and diplotype concordance in contrast to Astrolabe, Aldy, and Stargazer. This implies that ensemble genotyping followed by phenotyping, minimises incorrect predictions of individual *CYP2D6* diplotypes as well as metaboliser status for which the three algorithms could be prone if used singly.

The pros and cons of Astrolabe, Aldy, Stargazer, and the ensemble approach based on our analysis are summarised in Table 4.

Our approach had some limitations. Firstly, we used simulated data for most of the benchmarking especially for rare alleles and structural variants. Therefore, we did not capture the alterations in performance caused by some of the features present in real data

Table 4. Pros and cons of current *CYP2D6* genotyping algorithms.

Algorithm	Advantages	Disadvantages
Astrolabe	High accuracy for calling catalogued SNV-defined alleles.	Lower recall for *CYP2D6* CNVs and hybrid rearrangements compared to Aldy and Stargazer
	Supports hg19 and hg38.	Prone to ambiguous calls and miscalls if alleles/subalelles are not comprehensively defined
	Performs variant calling error correction.	Does not discriminate duplicated alleles in a sample with copy number gain (generically reports presence of a duplication)
	Easy to run (one-liner command).	Does not distinguish between duplication events and multiplication events
	Performs automated phenotype prediction.	
Aldy	High accuracy for calling catalogued SNV-defined alleles.	Aldy supports only hg19 as of v2.2.3
	Highest accuracy of the three tools in calling *CYP2D6* structural variations.	Prone to ambiguous calls and miscalls if alleles/subalelles are not comprehensively defined
	Easiest to run of the three tools as it requires only the BAM file on any system (even a normal laptop).	Prone to some erroneous calls when using default arguments due to sequencing noise
Stargazer	Supports various file inputs and imputation.	Low recall for rare alleles especially in heterozygous combinations
	Supports user-defined references for phasing and/or phased VCF input.	Affected by linkage disequilibrium as it has to do statistical reference-based phasing
	Provides tools for viewing metrics of key variants in a sample.	Prone to "no calls" for complex hybrid arrangements/combinations
	Outputs coverage plots for visually examining change-points for samples with CNVs.	Stargazer v1.0.7 supports only hg19
	Performs automated phenotype prediction.	Does not call subalelles as of v1.0.7
	Easy to run (one-liner command).	Has inconsistent reporting of alleles with uncertain function (reports background functionally annotated alleles for these cases as of v1.0.7)
Ensemble	Comparably high diplotype/haplotype concordance.	Difficult to resolve complete discrepancies between the genotypes of the three tools
	Resolves single tool deficiencies.	Prone to ambiguous calls especially for structural variations
		No automated pipeline as of now as the three algorithms are being updated on a regular basis and the reporting of alleles/subalelles is non-uniform

for these cases. However, publicly available reference datasets with validated *CYP2D6* diplotype calls are limited and they also have drawbacks as they do not cover a wide variety of *CYP2D6* allelic variation. The GeT-RM reference materials are more or less comprehensive; however, public WGS data is not yet available for all the samples or biobank scale cohorts. The different relative performances between the algorithms on simulated data and real data are attributed to the different distribution of alleles (i.e. fewer rare alleles in the real data). Our exhaustive simulation enabled us to test the genotyping algorithms on a wider array of *CYP2D6* variations. On the other hand, the small cohort of real world patients was very helpful for evaluating results from the algorithms against those obtained by using orthogonal techniques. The 75 GeT-RM samples are representative of 24 *CYP2D6* major star alleles (17 SNV-defined haplotypes plus 7 structural variants) compared to the 204 haplotypes (154 SNV-defined haplotypes plus 50 structural variants) represented in our simulated data.

We also did not compare the performance of the three tools across simulated or real NGS data from targeted custom-capture panels such as PGRNseq. These panels greatly reduce the cost of sequencing and enhance genotyping turnaround time. Further, we did not compare the influence of the aligner, BWA-MEM versus other aligners such as GSNAP[27] and NovoAlign (http://www.novocraft.com).

It is important to note that these algorithms face a major challenge of genotyping individuals with novel SNVs which may either be neutral or allele-defining (based on unpublished data). We did not test for this in this study as the reference materials used did not have novel star alleles. Ideally, for such cases, the two background star alleles would be called but the novel SNV(s) would be ambiguously assigned to either haplotype. Experimental validation using approaches such as Sanger sequencing, XL-PCR and/or long-read sequencing (e.g. PacBio, and Oxford Nanopore) would then be required to unequivocally determine the phase of these novel SNVs. Caution also needs to be taken in cases of complex and/or novel CNVs as they might be misreported by the algorithms. Nonetheless, combining short-read NGS (especially WGS) and tools such as Astrolabe, Aldy, and Stargazer (or their ensemble) is crucial to characterisation of pharmacogene variation in a cost-effective manner by reducing the number of cases that require the more expensive experimental genotyping approaches. Lastly, even though our comparison is focused on *CYP2D6*, we acknowledge that Astrolabe, Aldy, Stargazer, and PharmCAT (not included in the analysis) are extremely useful in genotyping other pharmacogenes for example *CYP2C9*, *CYP2C19*, *CYP2A6*, *CYP2B6*, *CYP3A4*, and *CYP3A5* as shown previously by others[15,16,20,28] (see also Supplementary Data Set 6).

Through this benchmarking study, we have examined the major differences as well as strengths and limitations of three algorithms that perform the challenging task of genotyping *CYP2D6* and other highly polymorphic pharmacogenes using short-read NGS data. Some of the inaccuracies reported were due to sequencing noise in the simulated and real data. This supports the notion that clinical-grade NGS data (preferably high coverage WGS) is required for effective use of these algorithms in the clinical setting. It is also clear that more *CYP2D6* pharmacogenetic studies especially in understudied populations would add to the completeness of the allele catalogue in PharmVar and in the allele definitions of these in silico genotyping algorithms hence enhancing their utility across populations.

METHODS

Ethics

Ethical approval was not required for this study as we used simulated and publicly available WGS data only.

Brief description for each algorithm

Detailed descriptions of all three algorithms are given in their original publications[15–17].

Briefly, the Astrolabe (v0.8.7.0) algorithm uses probability scoring to identify the most likely *CYP2D6* diplotype match based on variant positions and zygosity from input variant call format[29] (VCF) or gVCF files. Astrolabe also corrects for variant calling errors by using default or user-derived sensitivity and specificity of the variant calling approach. In addition, for WGS data, Astrolabe uses BAM[30] files for CNV calling by comparing the alignment depth of coverage ratios of the *CYP2D6* exonic regions to the coverage across various control regions along chromosome 22. Astrolabeś output includes *CYP2D6* diplotypes for each sample, special notes on structural variants detected, suballeles, and diplotype activity information.

Aldy (v2.2.3) differs from Astrolabe by using a combinatorial approach to reconstruct the sequence content in each copy of a target polymorphic gene based on data from a sample's SAM[31], BAM or CRAM[32] file. Aldy leverages integer programming solvers such as Gurobi[33], SCIP[34], and CBC[35] for timely solution of problem instances (copy number estimation, major and minor star allele prediction) that would otherwise take an extended period of time to solve. In addition, Aldy performs a coverage normalisation step, which corrects for errors when calling alleles and CNVs from non-uniform coverage sequence data generated from target panels such as PGRNseq.

The Stargazer (v1.0.7) pipeline retrieves variant information from VCF files generated using GATK-HaplotypeCaller[36], then performs statistical haplotype phasing using Beagle[37], and thereafter assigns *CYP2D6* diplotypes based on haplotype matches with a reference star allele table. The Stargazer pipeline also includes supplementary algorithms for phasing variants that are not present in the reference VCF (default in v1.0.7 is the 1000 genomes reference panel). Of note, Stargazer also accepts phased VCF files, chip-generated VCFs, and provides the option for imputation. In addition, for structural variant detection, Stargazer requires read depth information for the *CYP2D6* and *CYP2D7* regions as well as a control gene region (e.g. *EGFR*, *RYR1*, and *VDR*), generated using GATK-DepthOfCoverage[36]. The main output of Stargazer includes *CYP2D6* genotype data for each sample, interactive visual plots for CNVs, and predicted phenotype assignment based on the activity score system[38,39]. In case a user has an automation platform installed, such as the Sun Grid Engine (http://gridscheduler.sourceforge.net/htmlman/manuals.html), Stargazer can be run using the BAM file as input.

Simulation of benchmark datasets

In order to compare the performance of Aldy, Astrolabe, and Stargazer in calling a large variety of *CYP2D6* haplotypes, we created simulated datasets with known *CYP2D6* diplotypes for unrelated individuals, following the approach described by Numanagić et al.[19] with some modifications.

Maternal and paternal *CYP2D* locus sequences with respect to the human reference genome, GRCh37 (Chr22:42518000–42555000) were constructed separately for the simulation process. For cases with no allele-defining structural variants, GATK FastaAlternateReferenceMaker[36] was used to mutate the *CYP2D6* portion of the maternal and paternal *CYP2D* sequences by giving *CYP2D6* haplotype VCF files from PharmVar (version 4.0.4) as input. Haplotype sequences with *CYP2D6/2D7* intron and/or exon conversions were constructed in a similar way since PharmVar gives details of all the SNVs arising from these structural changes in the VCF files.

CYP2D6 gene deletions were generated by creating breakpoints within the REP6 and REP7 regions as described by Gaedigk et al.[40] and Steen et al.[41] previously. *CYP2D6* duplications were generated by concatenating *CYP2D6* sequences containing desired mutations. For haplotypes with more complex structural variations such as *CYP2D6 *61*, *CYP2D6*63*, and subvariants of *CYP2D6 *13*, we replaced reference *CYP2D6* and/or *CYP2D7* portions of the *CYP2D* locus with haplotype sequences deposited in the NCBI Nucleotide database by Kramer et al.[6], Black et al.[8], and Gaedigk et al.[7].

We simulated 101-bp paired-end reads (Illumina Hiseq2000) corresponding to maternal and paternal *CYP2D* haplotype sequences of each sample using the simNGS software (https://www.ebi.ac.uk/goldman-srv/simNGS/). simNGS replicates NGS-like noise intensity, base quality, and read error distributions by using runfiles trained from real sequencing experiments.

In order to account for the structural variant detection approaches used by Astrolabe and Stargazer, we simulated reads corresponding to CNV-neutral regions used by the algorithms and added them to the FASTQ files for each sample. The GRCh37 sequences (Chr22:19882000–19908000 and Chr22:44218000–44233000) contain control regions used by Astrolabe. For Stargazer, we used the *RYR1* gene locus (Chr19:38924340–39078204) as the control region. The CNV-neutral region that Aldy uses (Chr22:42547463–42548249) is covered by our simulation of the whole *CYP2D* locus.

Furthermore, to assess how various sequencing coverage affected Aldy, Astrolabe, and Stargazerś *CYP2D6* structural variant calling, we simulated read data with different coverage profiles that are frequently implemented in research and clinical settings (~30×, 60×, and 100×).

In total, we generated 4618 simulations for set 1, comprising only SNV-defined alleles. For set 2, where each sample had at least one haplotype with allele-defining *CYP2D6* strucural variant(s), we had 3 replicates of 121 samples at different coverages as mentioned above. We excluded some *CYP2D6* star alleles from our analysis, including **32*, **99*, **110–*139*, either because they were not defined by all the three tools or they were still under curation by PharmVar. However, undefined suballeles were included to check for their effect on the tools' performance.

Real data

Regarding the real data, we used 75 samples from the CDC's GeT-RM project for which WGS data is publicly available. These samples were originally collected as part of the 1000 genomes and HapMap projects and were sequenced on Illumina HiseqX platform (2 × 150 bp reads) (https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort#Pratt2016).

Variant calling

We aligned all our simulated read sets against the human reference genome (build 37) using BWA-MEM[30]. We used the b37 reference in our analysis because it is supported by all three algorithms (Table 1). The GeT-RM WGS data were already in BAM format (b37) on retrieval. We obtained gVCF files from BAM files using GATK-HaplotypeCaller followed by generation of VCF files with GATK GenomicsDB and GenotypeGVCFs[36].

CYP2D6 star allele calling

We performed *CYP2D6* star allele calling separately for Aldy, Astrolabe and Stargazer using default parameters, as well as parameters recommended by the authors in cases of debugging and CNV calling. BAM files were used as input for Aldy, whereas Astrolabe required both BAM and VCF files. For Stargazer, we first generated depth of coverage (GDF) files, required for calling structural variants, using GATK3 DepthofCoverage. The GDF files were then used as input together with VCF files for each run. We included reference samples without structural variants to account for Stargazer's intersample normalisation.

For the GeT-RM samples, we determined consensus genotypes (ensemble) from the separate genotype calls of Astrolabe, Aldy and Stargazer using the 2 from 3 rule. Truth calls for these samples were obtained from previous publications by others[23,24].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The real datasets used in this study are publicly available from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/data/view/PRJEB19931). Samples NA12878, NA12891, and N12892 can also be downloaded from the 1000 genomes project repository (https://www.internationalgenome.org). Metadata for all samples can be found at https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort#Pratt2016. The VCF files used to create hg19 consensus haplotype sequences for simulations are publicly available from the PharmVar database (https://www.pharmvar.org/gene/CYP2D6). All the fasta sequences used to generate *CYP2D6* CNV simulations are available at https://github.com/twesigomwedavid/CYP2D6-gt-algorithms.

CODE AVAILABILITY

The code used for this analysis is publicly available at https://github.com/twesigomwedavid/CYP2D6-gt-algorithms.

Received: 30 December 2019; Accepted: 20 May 2020;
Published online: 03 August 2020

REFERENCES

1. Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Klein, T. & Leeder, J. S. Prediction of CYP2D6 phenotype from genotype across world populations. *Genet. Med.* **19**, 69–76 (2017).
2. Zanger, U. M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.* **138**, 103–141 (2013).
3. Kimura, S., Umeno, M., Skoda, R. C., Meyer, U. A. & Gonzalez, F. J. The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic CYP2D6 gene, a related gene, and a pseudogene. *Am. J. Hum. Genet.* **45**, 889–904 (1989).
4. Drögemöller, B. I., Wright, G. E. B., Niehaus, D. J. H., Emsley, R. & Warnich, L. Next-generation sequencing of pharmacogenes: a critical analysis focusing on schizophrenia treatment. *Pharmacogenet. Genomics* **23**, 666–674 (2013).
5. Gaedigk, A. et al. The pharmacogene variation (PharmVar) consortium: Incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clin. Pharmacol. Ther.* **103**, 399–401 (2018).
6. Kramer, W. E. et al. CYP2D6: novel genomic structures and alleles. *Pharmacogenet. Genomics* **19**, 813–822 (2009).
7. Gaedigk, A. et al. CYP2D7-2D6 hybrid tandems: identification of novel CYP2D6 duplication arrangements and implications for phenotype prediction. *Pharmacogenomics* **11**, 43–53 (2010).
8. Black, J. L., Walker, D. L., O'Kane, D. J. & Harmandayan, M. Frequency of undetected CYP2D6 hybrid genes in clinical samples: impact on phenotype prediction. *Drug Metab. Dispos.* **40**, 111–119 (2012).
9. Nofziger, C. & Paulmichl, M. Accurately genotyping CYP2D6: not for the faint of heart. *Pharmacogenomics* **19**, 999–1002 (2018).
10. Gaedigk, A. Complexities of CYP2D6 gene analysis and interpretation. *Int. Rev. Psychiatry* **25**, 534–553 (2013).
11. Yang, Y., Botton, M. R., Scott, E. R. & Scott, S. A. Sequencing the CYP2D6 gene: from variant allele discovery to clinical pharmacogenetic testing. *Pharmacogenomics* **18**, 673–685 (2017).
12. Drögemöller, B. I. et al. CYP2D6 as a treatment decision aid for ER-positive non-metastatic breast cancer patients: a systematic review with accompanying clinical practice guidelines. *Breast Cancer Res. Treat.* **173**, 521–532 (2019).
13. Gordon, A. S. et al. PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenet. Genomics* **26**, 161–168 (2016).
14. Wright, G. E. B., Carleton, B., Hayden, M. R. & Ross, C. J. D. The global spectrum of protein-coding pharmacogenomic diversity. *Pharmacogenomics J.* **18**, 187–195 (2018).
15. Twist, G. P. et al. Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. *NPJ Genom. Med.* **1**, 15007 (2016).
16. Numanagić, I. et al. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nat. Commun.* **9**, 828 (2018).
17. Lee, S. et al. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet. Med.* **21**, 361–372 (2019).
18. Qiao, W. et al. The CYP2D6 VCF translator. *Pharmacogenomics J.* **17**, 301–303 (2017).
19. Numanagić, I. et al. Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data. *Bioinformatics* **31**, i27–34 (2015).
20. Sangkuhl, K. et al. Pharmacogenomics clinical annotation tool (PharmCAT). *Clin. Pharmacol. Ther.* **107**, 203–210 (2019).
21. Caudle, K. E. et al. Standardizing CYP2D6 genotype to phenotype translation: consensus recommendations from the clinical pharmacogenetics implementation consortium and dutch pharmacogenetics working group. *Clin. Transl. Sci.* **13**, 116–124 (2019).
22. Auton, A., Abecasis, G., Altshuler, D. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
23. Pratt, V. M. et al. Characterization of 137 genomic DNA reference materials for 28 pharmacogenetic genes: A GeT-RM collaborative project. *J. Mol. Diagn.* **18**, 109–123 (2016).
24. Gaedigk, A. et al. Characterization of reference materials for genetic testing of CYP2D6 alleles: a GeT-RM collaborative project. *J. Mol. Diagn.* **21**, 1034–1052 (2019).
25. Johansson, I. et al. Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proc. Natl Acad. Sci. USA* **90**, 11825–11829 (1993).
26. Nofziger, C. et al. PharmVar GeneFocus: CYP2D6. *Clin. Pharmacol. Ther.* **107**, 154–170 (2019).
27. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
28. Lee, S., Wheeler, M. M., Thummel, K. E. & Nickerson, D. A. Calling star alleles with stargazer in 28 pharmacogenes with whole genome sequences. *Clin. Pharm. Ther.* **106**, 1328–1337 (2019).
29. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
30. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
31. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
32. Fritz, M. H.-Y., Leinonen, R., Cochrane, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **21**, 734–740 (2011).
33. Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual* (2018).
34. Maher, S. J., et al. The SCIP Optimization Suite 4.0 (ZIB, Zuse Institute Berlin, 2017).
35. Perron, L., & Furnon, V. OR-Tools 7.2. Google (2019).
36. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
37. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
38. Gaedigk, A. et al. The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clin. Pharmacol. Ther.* **83**, 234–242 (2008).
39. Gaedigk, A., Dinh, J. C., Jeong, H., Prasad, B. & Leeder, J. S. Ten years' experience with the CYP2D6 activity score: A perspective on future investigations to improve clinical predictions for precision therapeutics. *J. Pers. Med.* **8**, 15 (2018).
40. Gaedigk, A., Blum, M., Gaedigk, R., Eichelbaum, M. & Meyer, U. A. Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. *Am. J. Hum. Genet.* **48**, 943–950 (1991).
41. Steen, V. M. et al. Detection of the poor metabolizer-associated CYP2D6(d) gene deletion allele by long-PCR technology. *Pharmacogenetics* **5**, 215–223 (1995).

ACKNOWLEDGEMENTS

We thank the developers of Astrolabe, Aldy, and Stargazer for technical assistance with running the algorithms. We acknowledge the Genetic Testing Reference Material Coordination Program (GeT-RM) for their contribution of characterised *CYP2D6* reference materials and whole-genome sequencing data for the corresponding samples. Special appreciation goes to Shaun Aron and Phelelani Mpangase for their support during the bioinformatics analysis at the Sydney Brenner Institute for Molecular Bioscience. We would also like to thank Annette Gross from the Clinical Pharmacology Modelling and Simulation Department in GlaxoSmithKline's Research and Development organisation for her perspectives on this work. This work was primarily funded through a grant by GlaxoSmithKline Research and Development Ltd to the Wits Health Consortium. G.S.K. was not responsible for the study design and the views and opinions expressed in this paper are not necessarily those of G.S.K. D.T and J.d.R. are partially funded by the South African National Research Foundation.

AUTHOR CONTRIBUTIONS

Study concept and design: D.T., S.H., Z.L, and J.d.R. Acquisition, analysis, or interpretation of the data: D.T., S.H., G.W., and B.D. Drafting of the manuscript: D.T., G.W., and B.D. Critical revision of the manuscript for important intellectual content: G.W., B.D., J.d.R., S.H., and Z.L. Obtained funding, administrative, technical or material support: S.H. Study supervision: S.H. and Z.L.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41525-020-0135-2.

Correspondence and requests for materials should be addressed to D.T. or S.H.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

StellarPGx: A Nextflow Pipeline for Calling Star Alleles in Cytochrome P450 Genes

David Twesigomwe[1,2,*], Britt I. Drögemöller[3], Galen E.B. Wright[4,5], Azra Siddiqui[6], Jorge da Rocha[1,2], Zané Lombard[2] and Scott Hazelhurst[1,6,*]

Bioinformatics pipelines for calling star alleles (haplotypes) in cytochrome P450 (CYP) genes are important for the implementation of precision medicine. Genotyping CYP genes using high throughput sequencing data is complicated, e.g., by being highly polymorphic, not to mention the structural variations especially in *CYP2D6*, *CYP2A6*, and *CYP2B6*. Genome graph-based variant detection approaches have been shown to be reliable for genotyping *HLA* alleles. However, their application to enhancing star allele calling in CYP genes has not been extensively explored. We present StellarPGx, a Nextflow pipeline for accurately genotyping CYP genes by combining genome graph-based variant detection, read coverage information from the original reference-based alignments, and combinatorial diplotype assignments. The implementation of StellarPGx using Nextflow facilitates its portability, reproducibility, and scalability on various user platforms. StellarPGx is currently able to genotype 12 important pharmacogenes belonging to the CYP1, 2, and 3 families. For purposes of validation, we use *CYP2D6* as a model gene owing to its high degree of polymorphisms (over 130 star alleles defined to date, including complex structural variants) and clinical importance. We applied StellarPGx and three existing callers to 109 whole genome sequenced samples for which the Genetic Testing Reference Material Coordination Program (GeT-RM) has recently provided consensus truth *CYP2D6* diplotypes. StellarPGx had the highest *CYP2D6* diplotype concordance (99%) with GeT-RM compared with Cyrius (98%), Aldy (82%), and Stargazer (84%). This exemplifies the high accuracy of StellarPGx and highlights its importance for both research and clinical pharmacogenomics applications. The StellarPGx pipeline is open-source and available from https://github.com/SBIMB/StellarPGx.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ As high throughput sequence data becomes increasingly available in clinical and research settings, bioinformatics tools for accurately calling star alleles in highly polymorphic pharmacogenes, via diverse approaches, are required to facilitate pharmacogenomics testing and allele characterization in various populations.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Our study addresses the need to implement a pipeline for accurate, scalable, and reproducible genotyping of cytochrome P450 (CYP) genes using high throughput sequence data. In addition, our pipeline performs automated phenotype prediction for well-characterized CYP genes such as *CYP2D6*.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ This study shows that StellarPGx can provide accurate, scalable, and reproducible CYP genotyping by combining pangenome-based variant detection, efficient combinatorial diplotype assignment, and read coverage information from high throughput sequence data.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ StellarPGx will facilitate CYP star allele calling in the clinical pharmacology and translational science communities. StellarPGx will also afford the ability to identify individuals with possible novel CYP alleles.

The human cytochrome P450 (CYP) gene family currently comprises 57 genes. Of these, 12 (belonging to the CYP1, 2, and 3 families) encode enzymes responsible for the bioactivation or elimination of over 80% of commonly prescribed medications. These CYP genes are also known to be highly polymorphic.[1] Accordingly, they account for the highest number of

[1]Sydney Brenner Institute for Molecular Bioscience (SBIMB), Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [2]Division of Human Genetics, National Health Laboratory Service, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [3]Department of Biochemistry and Medical Genetics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada; [4]Neuroscience Research Program, Kleysen Institute for Advanced Medicine, Winnipeg Health Sciences Centre and Max Rady College of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada; [5]Department of Pharmacology and Therapeutics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada; [6]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa. *Correspondence: David Twesigomwe (twesigomwedavid@gmail.com) and Scott Hazelhurst (scott.hazelhurst@wits.ac.za)

Received November 17, 2020; accepted January 5, 2021. doi:10.1002/cpt.2173

pharmacogenetic biomarkers in drug labels according to the US Food and Drug Administration (FDA).[2] Genotyping CYP genes and subsequent phenotype prediction is therefore a fundamental part of precision medicine approaches as CYP allele information is used for adjusting therapy, thus countering drug inefficacy and adverse drug reactions.[3–5]

CYP gene haplotypes—single nucleotide polymorphisms, small insertions and deletions (indels), and/or structural variations that are inherited together—are commonly denoted using the star (*) allele nomenclature to aid in interpretation.[6] We refer to single nucleotide polymorphisms and indels collectively as small nucleotide variations (SNVs) for the rest of this manuscript. The metabolizer activity of star alleles for most CYP genes ranges from nonfunctional to increased function, making it possible to predict individual phenotypes from diplotypes via activity score systems.[6–9] This makes accurate genotyping of CYP genes very useful and imperative. Notably, the Pharmacogene Variation Consortium (PharmVar)[10,11] and the Pharmacogenomics Knowledgebase (PharmGKB)[12] provide resourceful, curated catalogues of the haplotype information for CYP genes (and other pharmacogenes) as well as links to the related clinical annotations, respectively.

The ever-decreasing cost and increasing ubiquity of high-throughput sequencing (HTS) provide the opportunity for clinical-grade whole genome sequencing (WGS) and extensive characterization of CYP allele distribution especially in previously understudied populations. However, single molecule long-read technologies, which would make haplotype detection straightforward[13–16] are more expensive and less accessible compared with short-read technologies such as Illumina.[17] Notably, calling star alleles in CYP genes using short-read next-generation sequencing (NGS) data is challenging especially for genes which have neighboring pseudogenes (e.g., *CYP2D6*, *CYP2B6*, and *CYP2A6*) or isoforms (e.g., *CYP2C19* and *CYP2C9*) with high sequence similarity, hence causing a number of structural rearrangements, sequencing problems and subsequent read misalignments.[18,19]

The current state-of-the-art bioinformatics algorithms for calling star alleles in CYP and/or other pharmacogenes from short-read NGS data include Astrolabe (formerly Constellation),[20] Aldy,[21] Stargazer,[22,23] PharmCAT,[24] and Cyrius.[25] These algorithms mainly use binary alignment map (BAM) files and/or Variant Call Format (VCF) files as input to call SNV-defined alleles as well as structural variants. A notable commonality among these existing tools is that they all rely on variant detection based on reads mapped to a standard reference genome[26,27] (Genome Reference Consortium Human Build 37 (GRCh37) or GRCh38) as a fundamental step in their diplotype calling. Although this approach has facilitated variant calling from HTS data over the years, it has been shown to have limitations such as reference allele bias and read misalignment around indels.[28–30] Aligning reads to a reference genome without awareness of variation could, therefore, lead to erroneous variant calling and subsequently inaccurate star allele calling.

There is a growing paradigm shift towards using pangenomes to counter the limitations of reference-based alignments. A pangenome is a graph structure characterized by having a linear reference genome augmented with prior variant information such that each haplotype is represented as a path in the graph.[28–30] This facilitates more accurate read alignment as well as simultaneous genotyping of variants present in the graph.[31] Furthermore, novel variants can still be accurately called and used to update the graph. Given the high degree of polymorphisms and complex variation in most CYP genes, they are an excellent application for pangenome-based alignment and genotyping approaches. High genotyping accuracy, courtesy of genome graphs, has already been shown for similarly complex genomic regions such as the human leukocyte antigen (*HLA*) complex[32,33] and short tandem repeat regions.[34]

Here we present StellarPGx, a highly efficient hybrid pipeline that combines genome graph-based variant detection, read coverage information from the original HTS data alignments, and combinatorial diplotype assignment for genotyping CYP genes in a reproducible manner. StellarPGx is implemented using Nextflow,[35] an *in silico* workflow management system that facilitates parallelization, scalability, reproducibility, and portability of pipelines via Docker (https://www.docker.com) and Singularity[36] technologies. Docker and Singularity allow for the easy distribution of all the software dependencies for StellarPGx in a containerized system, thus overcoming reproducibility issues caused by the diversity in user computational frameworks.

We use *CYP2D6* as a model gene for validating StellarPGx using WGS data from 109 DNA samples that have been characterized as part of the Centers for Disease Control and Prevention's (CDC's) Genetic Testing Reference Materials Coordination Program (GeT-RM).[37] *CYP2D6* provides the ideal use case for StellarPGx as over 130 star alleles have been catalogued by PharmVar 10, including complex structural rearrangements with the neighboring nonfunctional paralog, *CYP2D7*. GeT-RM recently recharacterized 179 samples to resolve truth diplotypes for *CYP2D6* based on consensus results from various laboratories that used orthogonal genotyping techniques such as XL-PCR, Sanger sequencing, PacBio sequencing, and TaqMan copy number variation analysis.[37] One hundred and nine of the GeT-RM samples have publicly available short-read WGS data covering over 40 *CYP2D6* star alleles, which we used for the validation of StellarPGx and comparison with Aldy and Stargazer, which we had found to be the best-performing tools according to our previous benchmark,[38] as well as Cyrius 25 which was more recently developed.

METHODS

Implementation

We developed StellarPGx using Nextflow[35], a domain-specific language for creating scalable, highly parallelizable and portable pipelines with components that can be written in various programming languages. StellarPGx is containerized using Docker and Singularity for easy distribution of all the required software dependencies, thus facilitating workflow reproducibility on user systems. The major steps in the StellarPGx pipeline are summarized in **Figure 1** and described below using *CYP2D6* as a model gene. The StellarPGx version described here is currently able to identify *CYP2D6*1* through *CYP2D6*141*.

Input data and graph-based variant calling

StellarPGx takes BAM or CRAM files as input since they are the most common formats for storing and sharing WGS data that can be used for both SNV calling and copy number determination. StellarPGx supports alignments to both GRCh37 and GRCh38.

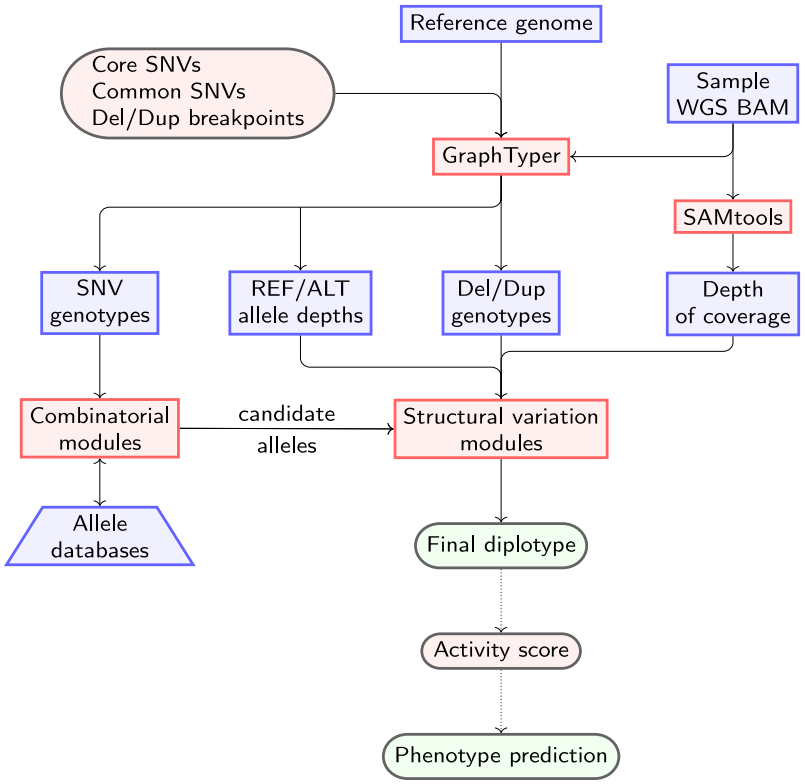


Figure 1 Overall summary of the StellarPGx workflow. The supported input in the version described here is a WGS (whole genome sequence) BAM (binary alignment map) file. Common and allele-defining SNVs (small nucleotide variations) as well as Del (gene deletion) and Dup (duplication) breakpoints are supplied in Variant Call Format for GraphTyper's pangenome construction. Copy number sentinel regions (target gene and control loci) are supplied in a BED (browser extensible data) file for read depth computation. StellarPGx's Nextflow processes then use the output from GraphTyper and SAMtools bedcov as well as input from allele/suballele combination databases to call sample diplotypes for a target cytochrome P450 (CYP) gene via Python 3.8 custom modules. Subsequent steps include determining the AS (activity score) based on the genotype followed by phenotype prediction depending on the AS group. These last two steps have not yet been implemented for all the supported CYP genes as consensus AS and/or phenotype assignment recommendations are available for only a few genes such as *CYP2D6*, *CYP2B6*, *CYP2C19*, and *CYP2C9*. ALT, alternate allele; REF, reference allele. [Colour figure can be viewed at wileyonlinelibrary.com]

Typical steps in pangenome-based variant detection (via the variation graph route rather than assembly) include genome graph construction using a reference genome and input variants (usually common variants in a population), graph indexing, alignment of sequence reads to the pangenome, and subsequent variant calling.[28] Rather than reinventing algorithms for these initial steps, StellarPGx uses GraphTyper.[33,39] GraphTyper is able to retrieve sequence reads from a defined region in a BAM/CRAM file and realign them to a variant-aware graph (pangenome), thus supporting SNV and structural variant genotyping based on the paths the reads align to and alignment coverage or breakpoint genotypes, respectively. StellarPGx uses GraphTyper to realign all reads from *CYP2D6* and *CYP2D7* to a genome graph with known *CYP2D6* variants, and simultaneously return variant genotypes as well as allele depth information in VCF format for each sample (see **Supplementary Materials S1** for an illustration of this process). Importantly, GraphTyper output also comprises novel sample variants.

The default prior variants included in StellarPGx for constructing the pangenome include all the allele-defining *CYP2D6* SNVs, as well as variants with a global frequency above 1% according to either the 1000 Genomes Project[40] or gnomAD[41] databases. For genotyping the *CYP2D6* full gene deletion and gene duplication via GraphTyper, StellarPGx uses a VCF with predefined imprecise breakpoints (see **Table S1**) for graph construction.

Star allele calling

StellarPGx's SNV-defined allele caller is written in Python 3.8 (https://docs.python.org/3.8/). It identifies known allele-defining variants from the VCF produced by GraphTyper and matches them to a plain-text StellarPGx database developed from PharmVar's star allele definitions.[10] The StellarPGx database contains all possible combinations of known SNV-defined alleles for *CYP2D6* and other supported CYP genes. For most cases, a combination of core SNVs corresponds to one diplotype. In such a scenario, this diplotype is considered to be the candidate diplotype in the sample. For the cases where the presence of a particular combination of core SNVs matches more than one candidate diplotype (typically two or three), we query a StellarPGx subdatabase with suballele combinations to find the candidate suballele-resolved combination that best explains all the SNV genotypes in the sample. This two-step process helps identify star alleles faster especially for cases where there is only one candidate diplotype. However, there are some candidate diplotypes which produce exactly the same variant combinations even when subvariants are considered. For example *CYP2D6*103.001/*3.001* and *CYP2D6*102.001/*3.002* produce a VCF with exactly the same variants as well as their zygosity (see **Supplementary Materials S1**). StellarPGx calls *CYP2D6*102/*3* or **103/*3* for a sample with such a combination. Such cases arise mostly due to alleles with a "limited" PharmVar level of evidence,[10]

meaning that their definitions are based on limited information such as exon-only sequencing as opposed to full gene resequencing.

The next step in the star allele calling process is discriminating structural variant alleles, if present. Due to the many *CYP2D6* structural variant rearrangements,[10] we integrate read coverage data from the original reference-based WGS BAM/CRAM file as well to support calling of complex *CYP2D6* structural rearrangements with *CYP2D7*. We use SAMtools[42] to compute the depth of coverage for predefined *CYP2D6*, *CYP2D7*, and control gene regions as well as exons/introns affected by known hybrid rearrangements. For example, exon 9 is a priority as there are a number of *CYP2D6-2D7* alleles with an exon 9 conversion including *CYP2D6*36*, **57*, and **83*.[10] Details for the GRCh37 and GRCh38 regions used in StellarPGx's structural variant calling can be found in the StellarPGx project repository (https://github.com/SBIMB/StellarPGx). Vitamin D Receptor (*VDR*) and Epidermal Growth Factor Receptor (*EGFR*) genes have been shown to be effective as control genes in star allele calling[23] and they have relatively low copy number variation frequencies;[43–45] hence we use the average coverage of both genes as a control for copy number. The implementation of StellarPGx in Nextflow allows for this depth-of-coverage computation to be performed in parallel with the aforementioned pangenome-based variant calling.

We determine the total copy number of *CYP2D6* alleles in the sample by comparing the read coverage for *CYP2D6* and its flanking regions with the read coverage for the control loci. If a copy number of 2 is ascertained, StellarPGx reports the candidate SNV diplotypes as the final result except for cases where we check for presence of possible singleton hybrids. For example, we use individual algorithms to check whether *CYP2D6*36* (exon 9 conversion) is present instead of the initially called *CYP2D6*10*. If a copy number of 1 is detected, StellarPGx replaces the homozygous diplotype assignments from the SNV information with a *CYP2D6 *5/**<allele>assignment (**5* represents the *CYP2D6* gene deletion allele). For example, a sample with *CYP2D6*5/*17* would initially be genotyped as *CYP2D6*17/*17* based on SNVs alone. After running the SV caller algorithm, StellarPGx would then report the final diplotype as *CYP2D6*5/*17*. The detection of copy number 0 would default to a *CYP2D6*5/*5* diplotype call.

On the other hand, if a copy number of 3 or more is detected, StellarPGx uses the allele depth ratios (from GraphTyper VCF output) for each core SNV to decipher which of the two alleles in a sample is duplicated/multiplicated. For duplications involving tandem rearrangements, StellarPGx runs hybrid caller modules to discriminate the hybrid copies based on information from PharmVar structural variant definitions.[10] For example, let us consider the *CYP2D6*68+*4* tandem (found frequently in Europeans) which comprises the nonfunctional **4* allele and the *CYP2D6-2D7* hybrid, i.e., *CYP2D6**68 (intron 1 switch region). StellarPGx ascertains the *CYP2D6*68+*4* tandem in a sample with the *CYP2D6*41/*68+*4* diplotype by comparing the coverage from the promoter region to intron 1 vs. the coverage from intron 1 to exon 9 as well as the allele depth (and zygosity) for 100C > T (found in *CYP2D6**68 and most *CYP2D6*4* alleles) vs. that for 1847G > A (*4) (see **Figure 2b**). Other significantly frequent tandems include *CYP2D6*36+*10, *36×2*, and **36×2+*10*, which can be discriminated from *CYP2D6*10×2* and **10×3* by interrogating the *CYP2D6* exon 9 coverage (see **Figure 2c**). In some cases (such as *CYP2D6*13* hybrid alleles—see **Figure 2d**) the coverage of *CYP2D7* regions is also considered to enhance StellarPGx's recall. More details about the modules for calling the wide array of CYP structural variants (especially for *CYP2D6*) can be found in the StellarPGx project repository (https://github.com/SBIMB/StellarPGx).

StellarPGx output

The StellarPGx output file contains the final diplotype call and phenotype assignment (where applicable) of the target CYP gene as well as a breakdown of the variants and copy number information used for diplotype assignment. Other candidate solutions that were considered by the caller (if applicable) are also provided as this information could be useful especially when performing ensemble star allele calling (using at least three tools to enhance accuracy)[38]. Example output files are provided in **Supplementary Materials S1**.

For cases where StellarPGx detects a novel allele or suballele, it will indicate "Possible novel major allele or suballele present: Interpret with caution; Experimental validation and expert review through PharmVar is recommended" in the output (see **Supplementary Materials S1**). Novel allele discovery is more or less straightforward for samples with novel combinations of existing core variants. However, detecting novel alleles when completely novel core variants are present in a sample is more challenging. The current strategy is to include novel (i.e., not yet in the PharmVar database) missense, stop loss, stop gain, and frameshift CYP variants in the StellarPGx database so these can be flagged by StellarPGx when they appear in the GraphTyper-generated VCF. These variants include SNVs identified from large-scale genomics projects such as the 1000 Genomes Project[40] and gnomAD[41], as well as possible missense variants that could occur in the target CYP genes even though they haven't been identified in previous studies. However, we are yet to include support for calling novel structural variants. The structural variants that are currently supported by StellarPGx are those defined by PharmVar as per the v4.2.4 update.[10] We strongly recommend that users liaise with PharmVar[10] for information about experimental validation of possible novel star alleles and/or their designation if any of these cases arise in clinical or research settings.

Validation

In order to validate the StellarPGx pipeline, we performed *CYP2D6* star allele calling on publicly available high coverage WGS data for 109 ethnically diverse reference samples from the GeT-RM project. This includes 75 samples from the previous GeT-RM work[46] and 34 of the 42 new samples added in the latest GeT-RM study[37], which was focused on recharacterization of *CYP2D6* alleles for 179 DNA samples (not all have WGS data publicly available) derived from Coriell cell lines. (See **Table S2** for details about the data sources.)

We compared *CYP2D6* star allele calls and phenotype assignments by StellarPGx (version described here covers *CYP2D6*1-*141*) with the GeT-RM consensus reference materials (cover *CYP2D6*1-*114*), Aldy v2.2.6 (covers *CYP2D6*1-*114*), Stargazer v1.0.8 (covers *CYP2D6*1-*138*), and Cyrius v1.1 (covers *CYP2D6*1-*139*). For Aldy and Stargazer, we performed read realignment (*CYP2D* and control regions) for 34 of the samples from GRCh38 to GRCh37 using BWA-MEM.[26] We used the *VDR* gene as the control region for the Stargazer run. The control regions for Aldy and Cyrius are predetermined within their algorithms.

Furthermore, we genotyped 11 other CYP genes (including *CYP2A6*, *CYP2B6*, *CYP2C19*, *CYP2C9*, *CYP2C8*, *CYP3A4*, *CYP3A5*, *CYP2A1*, *CYP1A2*, *CYP2E1*, and *CYP4F2*) using 75 WGS samples corresponding to reference materials from the GeT-RM study by Pratt *et al.*,[46] to demonstrate the utility of StellarPGx in characterizing the diverse CYP allele landscape. However, for some genes, not all the samples had corresponding GeT-RM consensus calls reported. Therefore, StellarPGx's concordance was determined using the samples for which the GeT-RM consensus calls were available.

RESULTS

Comparison of StellarPGx with other callers based on *CYP2D6* diplotyping

We used published consensus *CYP2D6* diplotypes for the 109 WGS samples corresponding to the GeT-RM reference materials as the truth diplotype calls 38 except for sample NA18519. GeT-RM reported the consensus *CYP2D6* call for NA18519 as **1/*29*. However, this sample's reference diplotype was based on

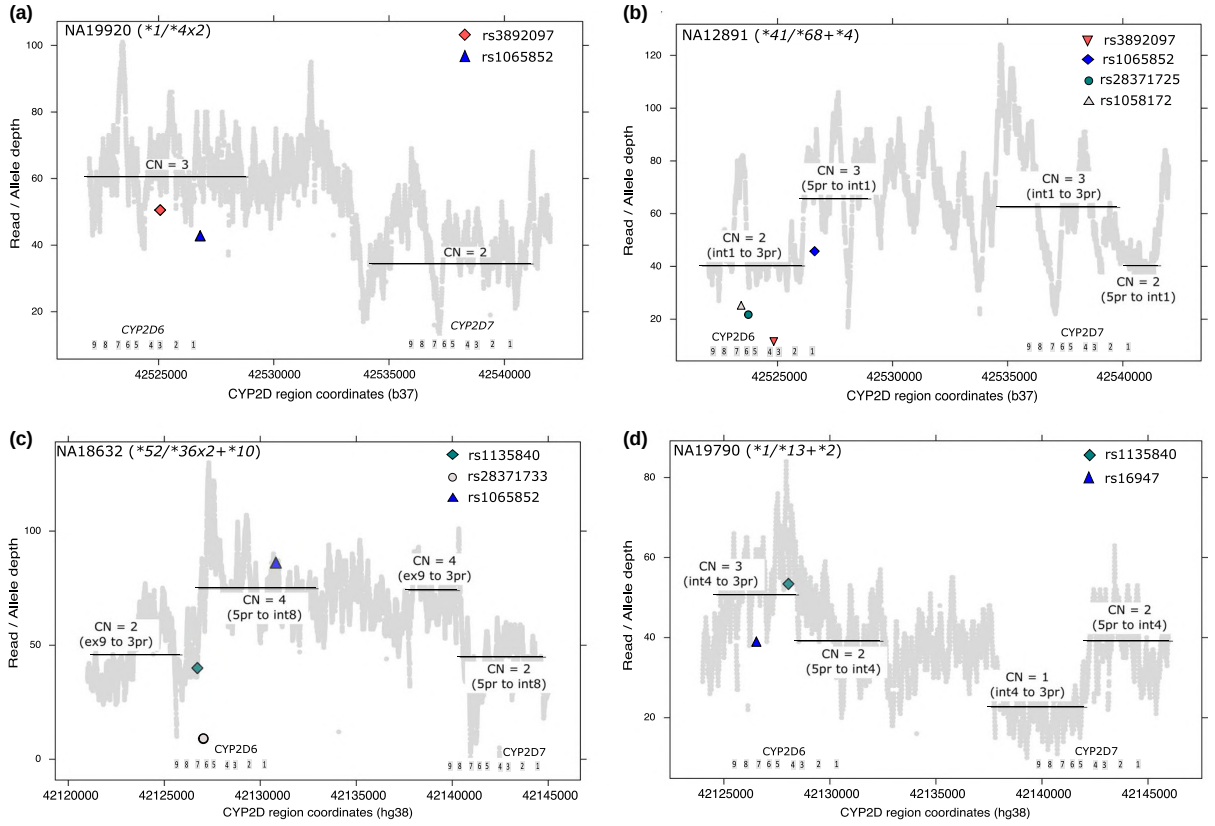


Figure 2 Examples of *CYP2D6* structural variants detected by StellarPGx in the WGS validation data set corresponding to the GeT-RM reference materials. (**a**) Shows the *CYP2D* region coverage in an admixed American sample NA19920 with the *CYP2D6*1/*4×2* diplotype. Note the higher *CYP2D6* copy number (CN = 3) compared with *CYP2D7* (CN = 2) and the comparable allele depth for the core variants (rs3892097 and rs1065852). (**b**) Shows a European sample NA12891 with the *CYP2D6*41/*68+*4* diplotype. Note the hybrid read distribution due to the *68 allele and the intron 1 switch region. Also, the allele depth for rs1065852 is much higher than that for rs3892097. (**c**) Shows an East Asian sample NA18632 with the complex *CYP2D6*52/*36×2+*10* diplotype. Note the exon 9 conversion showed by diminished coverage for the *CYP2D6* exon 9 and elevated coverage for *CYP2D7* exon 9. (**d**) Shows sample NA19790 with the complex *CYP2D6*1/*13+*2* diplotype. Note the intron 4 switch region for the *CYP2D6*13* hybrid allele. CYP, cytochrome P450; ex, exon; int, intron; WGS, whole genome sequence; 3pr, 3-prime region; 5pr, 5-prime region. [Colour figure can be viewed at wileyonlinelibrary.com]

targeted genotype panel results only,[38] thus probably missing the **106* defining variant (3878G>A), which appears in the WGS data. Therefore, we considered *CYP2D6*29/*106* as the truth call for NA18519 as did the Cyrius authors.[25] The assays used to generate the GeT-RM truth calls are described in detail by Gaedigk *et al*.[37]

StellarPGx's *CYP2D6* star allele calls were highly concordant (99%) with the GeT-RM reference materials and comparable to Cyrius calls (**Table 1**). Notably, StellarPGx was able to correctly genotype sample NA19908 (*CYP2D6*1/*46*— normal metabolizer phenotype) as the pipeline queried multiple suballele combinations, unlike Cyrius which ambiguously reported *CYP2D6*1/*46* or **1/*43* (indeterminate phenotype) for this sample. In comparison, Aldy and Stargazer diplotype calls had lower concordance with GeT-RM calls (**Table 1**).

Furthermore, StellarPGx's phenotype assignments were 100% concordant with those corresponding to the GeT-RM reference materials. Cyrius had comparable phenotype concordance (99%), while Aldy (92%) and Stargazer (90%) had lower phenotype assignment accuracy. Activity scores and metabolizer phenotypes were assigned according to consensus recommendations by Caudle *et al*.[6] Detailed diplotype calls and corresponding phenotype assignments for each sample by each tool are provided in **Table S3**.

Sample NA18545 (*CYP2D6*5/*36×2+*10×2*) was particularly challenging for all four tools as the **5* allele could not be detected in the presence of the *CYP2D6*36×2+*10×2* allele on the other chromosome. This sample was not genotyped by Stargazer while StellarPGx, Cyrius, and Aldy genotyped it as *CYP2D6*36+*10/*36+*10*. Both diplotypes, however, translate to an intermediate metabolizer phenotype when applying the Clinical Pharmacogenetics Implementation Consortium (CPIC)–recommended translation method 6. StellarPGx consistently called the other diplotypes comprising *CYP2D6* hybrid rearrangements (**Figure 2**).

Notably, some indel positions in complex gene regions had to be shifted in our allele databases to call alleles such as *CYP2D6*40* and *CYP2D6*21*. This shift in indel position is also implemented by the other allele callers.

Table 1 Summary of the *CYP2D6* diplotype concordance for StellarPGx in comparison with other tools

Allele caller	Concordant samples	Concordance (%)
StellarPGx	108	99
Cyrius	107	98
Stargazer	92	84
Aldy	89	82

Summary of the *CYP2D6* diplotype concordance for StellarPGx in comparison with other tools with respect to the 109 GeT-RM WGS samples used in the study.
CYP, cytochrome P450; GeT-RM, Genetic Testing Reference Material Coordination Program; WGS, whole genome sequence.

StellarPGx's performance on a larger *CYP2D6* diplotype sample space

Given that the number of *CYP2D6* alleles represented in the GeT-RM reference materials was rather limiting, we further tested the ability of StellarPGx's combinatorial diplotype assignment to resolve 35,778 diplotype simulations. This includes all combinations of 267 SNV-defined haplotypes covering *CYP2D6*1.001-*140.001* (see **Table S4**) from PharmVar v4.2.4 corresponding to GRCh37. StellarPGx correctly resolved 35,764 diplotypes and reported ambiguous calls for 14 cases. Examples of the ambiguous cases encountered include *CYP2D6*132.001/*9.001* (called as *CYP2D6*10/*115* or **132/*9*), *CYP2D6*1.001/*6.003* (called as *CYP2D6*1/*6* or **39/*6*), and *CYP2D6*2.004/*6.002* (called as *CYP2D6*2/*6* or **34/*6*). See **Supplementary Materials S1** and **Table S4** for more details on the simulation analysis.

StellarPGx's performance on 11 additional CYP genes

StellarPGx's concordance with GeT-RM calls for *CYP2A6*, *CYP2B6*, *CYP2C19*, *CYP2C9*, *CYP2C8*, *CYP3A4*, *CYP3A5*, *CYP2A1*, *CYP1A2*, *CYP2E1*, and *CYP4F2* diplotypes is summarized in **Table 2**. Briefly, StellarPGx had 100% concordance with GeT-RM diplotype calls 48 for *CYP1A2*, *CYP3A4*, *CYP3A5*, and two *CYP2C* cluster genes, i.e., *CYP2C9* and *CYP2C8*. We also obtained highly comparable diplotypes with GeT-RM calls for *CYP4F2* (96%), *CYP2C19* (97%), and *CYP2A6* (88%). On the other hand, genotyping *CYP1A1*, *CYP2E1*, and *CYP2B6* was more challenging as the StellarPGx concordance with GeT-RM calls with respect to these genes was 78%, 63%, and 66%, respectively. For a detailed explanation of the discrepancies in the called diplotypes for *CYP1A1*, *CYP2E1*, and *CYP2B6*, see Discussion section. Detailed star allele calls for each gene/sample are provided in **Table S5**.

Of note, we identified samples with possible novel alleles for some CYP genes. For example, samples NA07056, NA07029, and NA18564 were genotyped as *CYP4F2*3/*3* by GeT-RM, but StellarPGx flagged that a possible novel *CYP4F2* allele was present as rs2108622 (*CYP4F2*3*) and rs3093105 (*CYP4F2*2*) were deduced to be on the same chromosome in each of these samples.

Run time comparisons

We compared the run time of StellarPGx against Aldy, Stargazer, and Cyrius for *CYP2D6* star allele calling on a cluster of 4 computers, each with dual Intel Silver 4114 processors and 132GB of RAM, and running CentOS 7.5 (https://www.centos.org). Each task could use 20 GB memory and 10 parallel threads. For a single WGS sample run, Aldy was the fastest tool with a run time of 7 seconds while StellarPGx (58 seconds), Stargazer (91 seconds), and Cyrius (140 seconds) were significantly slower. However, for a multi-sample run (75 WGS samples aligned to build 37), StellarPGx had the fastest run time (4 minutes 7 seconds) followed by Aldy (6 minutes) and Stargazer (9 minutes 51 seconds), while Cyrius was significantly slower (57 minutes 49 seconds when using a single manifest file). StellarPGx was able to obtain the fastest run time for the multisample run owing to the efficient parallelization facilitated by its Nextflow implementation.

DISCUSSION

Accurately calling star alleles in CYP genes is integral to genotype-guided therapy for a wide array of drugs, and characterizing pharmacogenomic variation at the population level. Reproducible bioinformatics workflows are crucial in tackling the challenging task of genotyping this class of pharmacogenes using targeted and/or WGS data from the widely used short-read NGS technologies. Existing tools, notably Astrolabe, Aldy, Stargazer, and Cyrius have provided short-read-based genotyping solutions for CYP and/or other pharmacogenes in recent years. However, more tools with diverse approaches are still needed especially to support high confidence ensemble CYP genotyping that we propose in our previous benchmarking study.[38] As long read technologies such as PacBio[13,14] and Oxford Nanopore[15,16] become more widely used, adapting star allele callers to the analysis of targeted and/or WGS long-read data will presumably be easier, but nonetheless important.

We have described StellarPGx, a Nextflow pipeline for determining diplotypes in CYP genes from short-read HTS data, and

Table 2 Summary of the diplotype concordance for StellarPGx with respect to 11 additional *CYP* genes

Gene	Number of samples	Concordant samples	Concordance (%)
CYP2C8	74	74	100
CYP2C9	74	74	100
CYP3A4	74	74	100
CYP3A5	74	74	100
CYP1A2	67	67	100
CYP2C19	74	72	97
CYP4F2	67	64	96
CYP2A6	67	59	88
CYP1A1	67	52	78
CYP2B6	74	49	66
CYP2E1	67	42	63

Summary of the diplotype concordance for StellarPGx with respect to 11 additional *CYP* genes in comparison with the GeT-RM reference materials. The sample number ranges from 67 to 75 (all aligned to build 37) as not all the 109 aforementioned WGS samples had publicly available truth calls from GeT-RM.
CYP, cytochrome P450; GeT-RM, Genetic Testing Reference Material Coordination Program; WGS, whole genome sequence.

demonstrated its analytic accuracy using *CYP2D6* as a model CYP gene. StellarPGx uses a hybrid approach involving genome graph-based variant calling, read coverage information from the original linear alignments, and combinatorial star allele assignment. StellarPGx's *CYP2D6* allele calling performance is comparable to the recently developed Cyrius pipeline, and is superior to Aldy and Stargazer based on the analysis of WGS data corresponding to the GeT-RM reference materials.

Furthermore, StellarPGx's phenotype assignments were more concordant with the GeT-RM reference materials compared with any of the other allele callers (**Table S3**). The effect of genotyping accuracy on phenotype assignment, however, depends on the sample diplotype. For example, sample NA18565 (*CYP2D6*10/*36×2*) was correctly assigned an intermediate metabolizer phenotype by all four allele callers, even though Aldy and Stargazer incorrectly called it as *CYP2D6*10/*36+*10*. Conversely, sample NA18524 (*CYP2D6*1/*36×2+*10*, normal metabolizer) was incorrectly assigned an intermediate metabolizer phenotype by Stargazer, which genotyped it as (*CYP2D6*36 + 10/*36+*10*).

Though comparable in genotyping performance, StellarPGx offers notable advantages over Cyrius. Firstly, StellarPGx is able to call diplotypes in up to 12 CYP genes while Cyrius is currently only available for *CYP2D6*. Secondly, owing to its implementation in Nextflow and containerization with both Docker and Singularity, StellarPGx installs easily across various computational platforms and affords high reproducibility and scalability. In the same vein, StellarPGx calls star alleles in a shorter time compared with Cyrius especially for large cohorts. In addition, users can easily perform modifications to the StellarPGx pipeline by including custom Nextflow processes for performing different steps. One such modification would be to run a tool like Aldy in a parallel process so as to obtain consensus genotype calls for the same CYP gene(s) in multiple samples. Furthermore, the use of pangenome-based variant calling in the initial process of the StellarPGx pipeline enhances allele-calling confidence. In addition, StellarPGx is able to detect possible occurrences of novel alleles by flagging novel combinations of existing core SNVs and also interrogating whether novel core SNVs (especially missense, frameshift, stop loss, and stop gain variants) are present in a sample.

As shown in our simulations, StellarPGx is currently not able to resolve diplotypes with exactly similar variant combinations (see **Supplementary Materials S1**), thus resulting in ambiguous diplotype assignments for such cases. This limitation is also common to other allele callers including Astrolabe,[20] Aldy,[21] and Cyrius.[25] Future updates to the StellarPGx database with alleles upgraded from limited to moderate or definitive level of evidence by PharmVar will likely mitigate this challenge. Another main limitation is having to perform structural variant detection by complementing GraphTyper's structural variant genotypes with coverage information from original linear alignments. Even though genome graph-based variant detection has proved to be excellent for genotyping SNVs, full gene deletion, and full gene duplications, we observed a number of false positive and false negative genotypes for the hybrid rearrangements in the model gene (*CYP2D6*) and challenges in including them in the pangenome construction, hence resorting to the hybrid allele calling approach. Additionally, the realignment by GraphTyper to the pangenome is done for reads already aligned to a genomic region of interest in the original linear reference alignment, which could introduce some reference-based alignment bias.

Further, StellarPGx's diplotype concordance was lower for *CYP1A1*, *CYP2E1*, and *CYP2B6* compared with other supported CYP genes. For *CYP2B6* in particular, StellarPGx had a high number of ambiguous diplotype calls (20 samples) with respect to the data set used for our validation. For example, samples HG00436, NA06991, NA07029, NA07055, and NA10847 (genotyped as *CYP2B6*1/*6* by GeT-RM as **6* is considerably more frequent than **4*) were genotyped as *CYP2B6*1/*6* or **4/*9* by StellarPGx since both diplotypes are equally likely given the allele defining variant combinations in the samples. For *CYP1A1* and *CYP2E1*, one possible explanation for the inconsistent diplotype calls is that the number of variants tested by some of the genotyping panels used by GeT-RM was rather limiting. For example, sample NA18980 was called as *CYP1A1*1/*2* by GeT-RM, while StellarPGx called it as *CYP1A1*13/*2B* since we were able to detect the *CYP1A1*13* defining variant rs4646422. Notably, for *CYP2E1*, only two panels were used for generating consensus reference *CYP2E1* genotypes by GeT-RM. Indeed most of the discrepant *CYP2E1* diplotypes between StellarPGx and GeT-RM arose in samples that did not have definitive consensus calls by GeT-RM. For example, samples HG00436 and NA18980 were homozygous for both *CYP2E1*5B* (rs3813867, rs2031920) and *CYP2E1*7* (rs2070673) defining variants. GeT-RM genotyped them as *CYP2E1(*7/*7)*, indicating uncertainty in the final diplotype, while StellarPGx had "no calls" for these samples and indicated possible presence of novel alleles. Similarly, sample NA18992 was genotyped by GeT-RM as *CYP2E1(*7)/*7* while StellarPGx detected the potentially function-altering rs199856651 in this sample. Furthermore, possible novel alleles were also identified for *CYP4F2* and *CYP2A6* in the validation data set (see **Table S5**). These findings indicate that more comprehensively characterized reference materials are needed for all core CYP genes as has been undertaken for *CYP2D6* recently.[37] Additionally, updates to PharmVar's star allele definitions for genes such as *CYP2E1* and *CYP2A6* will allow for allele definitions to better reflect variation in their loci.

Future developments of StellarPGx include adding support for more CYP genes as well as related pharmacogenes such as *POR*. We also intend to make StellarPGx a fully pangenome-based star allele calling pipeline as structural variant genotyping using genome graph approaches improves. Additionally, adding support for long-read sequence data and targeted NGS data from panels such as PGRNseq[47] would enhance the utility of StellarPGx. However, analyzing whole exome sequence data is still a major challenge as intronic SNVs and copy number change-points are also important for StellarPGx's star allele calling.

For short-read data, we recommend running StellarPGx using high coverage (at least 25×) WGS data especially when genotyping *CYP2D6*, *CYP2A6*, *CYP2B6*, and *CYP2E1*. Further efforts aim to keep the automated phenotype prediction up to date as activity score systems and phenotype assignment recommendations are refined by clinical pharmacogenomics consortia.

We plan to continue enhancing StellarPGx's ability to detect novel alleles as well as complex population-specific alleles as community usage increases, especially from studies using NGS data from underrepresented populations. StellarPGx's allele databases will also be updated on a regular basis to match the updates to reference databases such as PharmVar 10, especially as comprehensive allele characterization of genes such as *CYP2E1* and *CYP2A6* becomes available.

We recommend running StellarPGx in a multinode cluster-environment using the prebuilt Singularity image. However, for single sample cases, the pipeline can also be run on a personal computer (MacOS or Linux) preferably enhanced with external disk storage. StellarPGx is freely available from https://github.com/SBIMB/StellarPGx under the MIT License.

SUPPORTING INFORMATION

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

ACKNOWLEDGMENTS

We acknowledge the Centers for Disease Control and Prevention's (CDC's) Genetic Testing Reference Material Coordination Program (GeT-RM) for providing CYP2D6 reference materials for validation of pharmacogene genotyping approaches. We also acknowledge Illumina for generating 75 of the WGS samples (including the Polaris pharmacogenomics cohort) used in the validation of StellarPGx. Samples NA19790, NA10860, and HG00423 were generated from Coriell cell lines at the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository. Thirty-one of the high coverage 1000 Genomes Project data sets were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

FUNDING

This work was primarily funded through a grant by the GlaxoSmithKline (GSK) Research and Development Ltd to the Wits Health Consortium. GSK was not responsible for the study design, and the views and opinions expressed in this paper are not necessarily those of GSK. D.T. (SFH210111581508) and J.d.R. (SFH170626244782) are supported in part by the South African National Research Foundation.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

D.T., B.I.D., G.E.B.W., A.S., J.d.R., Z.L., and S.H. wrote the manuscript. D.T., B.I.D., G.E.B.W., J.d.R., Z.L., and S.H. designed the research. D.T. and S.H. performed the research. D.T. analyzed the data.

1. Zanger, U.M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.* **138**, 103–141 (2013).
2. US Food and Drug Administration. Table of pharmacogenomic biomarkers in drug labeling <https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm> (2020). Accessed September 20, 2020.
3. Goetz, M.P. *et al*. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6 and tamoxifen therapy. *Clin. Pharmacol. Ther.* **103**, 770–777 (2018).
4. Rajman, I., Knapp, L., Morgan, T. & Masimirembwa, C. African genetic diversity: implications for cytochrome P450-mediated drug metabolism and drug development. *EBioMedicine* **17**, 67–74 (2017).
5. Zhou, Y., Ingelman-Sundberg, M. & Lauschke, V.M. Worldwide distribution of cytochrome P450 alleles: a meta-analysis of population-scale sequencing projects. *Clin. Pharmacol. Ther.* **102**, 688–700 (2017).
6. Caudle, K.E. *et al*. Standardizing CYP2D6 genotype to phenotype translation: consensus recommendations from the Clinical Pharmacogenetics Implementation Consortium and Dutch Pharmacogenetics Working Group. *Clin. Transl. Sci.* **13**, 116–124 (2020).
7. Gaedigk, A., Dinh, J.C., Jeong, H., Prasad, B. & Leeder, J.S. Ten years' experience with the CYP2D6 activity score: a perspective on future investigations to improve clinical predictions for precision therapeutics. *J. Pers. Med.* **8**, 15 (2018).
8. Gaedigk, A., Simon, S.D., Pearce, R.E., Bradford, L.D., Kennedy, M.J. & Leeder, J.S. The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clin. Pharmacol. Ther.* **83**, 234–242 (2008).
9. Zubiaur, P. *et al*. Influence of CYP2B6 activity score on the pharmacokinetics and safety of single dose efavirenz in healthy volunteers. *Pharmacogenomics J.* **20**, 235–245 (2020).
10. Gaedigk, A. *et al*. The Pharmacogene Variation (PharmVar) Consortium: incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clin. Pharmacol. Ther.* **103**, 399–401 (2018).
11. Gaedigk, A. *et al*. The evolution of PharmVar. *Clin. Pharmacol. Ther.* **105**, 29–32 (2019).
12. Whirl-Carrillo, M. *et al*. Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).
13. Buermans, H.P.J. *et al*. Flexible and scalable full-Length CYP2D6 long amplicon PacBio sequencing. *Hum. Mutat.* **38**, 310–316 (2017).
14. Qiao, W. *et al*. Long-read Single Molecule Real-Time full gene sequencing of cytochrome P450–2D6. *Hum. Mutat.* **37**, 315–323 (2016).
15. Liau, Y., Maggo, S., Miller, A.L., Pearson, J.F., Kennedy, M.A. & Cree, S.L. Nanopore sequencing of the pharmacogene CYP2D6 allows simultaneous haplotyping and detection of duplications. *Pharmacogenomics* **20**, 1033–1047 (2019).
16. Ammar, R., Paton, T.A., Torti, D., Shlien, A. & Bader, G.D. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Res.* **4**, 17 (2015).
17. Bentley, D. *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
18. Yang, Y., Botton, M.R., Scott, E.R. & Scott, S.A. Sequencing the CYP2D6 gene: from variant allele discovery to clinical pharmacogenetic testing. *Pharmacogenomics* **18**, 673–685 (2017).
19. Nofziger, C. & Paulmichl, M. Accurately genotyping CYP2D6: not for the faint of heart. *Pharmacogenomics* **19**, 999–1002 (2018).
20. Twist, G.P. *et al*. Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, *CYP2D6*, from whole-genome sequences. *NPJ Genom. Med.* **1**, 15007 (2016).
21. Numanagić, I. *et al*. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nat. Commun.* **9**, 828 (2018).
22. Lee, S.-B. *et al*. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet. Med.* **21**, 361–372 (2019).
23. Lee, S.-B., Wheeler, M.M., Thummel, K.E. & Nickerson, D.A. Calling star alleles with Stargazer in 28 pharmacogenes with whole genome sequences. *Clin. Pharmacol. Ther.* **106**, 1328–1337 (2019).
24. Sangkuhl, K. *et al*. Pharmacogenomics clinical annotation tool (PharmCAT). *Clin. Pharmacol. Ther.* **107**, 203–210 (2020).
25. Chen, X. *et al*. Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data. *Pharmacogenomics J.* (2021). https://doi.org/10.1038/s41397-020-00205-5
26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

27. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
28. Garrison, E. *et al*. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
29. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
30. Ambler, J.M., Mulaudzi, S. & Mulder, N. GenGraph: a python module for the simple generation and manipulation of genome graphs. *BMC Bioinformatics* **20**, 519 (2019).
31. Rakocevic, G. *et al*. Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).
32. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
33. Eggertsson, H.P. *et al*. Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
34. Dolzhenko, E. *et al*. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
35. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. & Notredame, C. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
36. Kurtzer, G.M., Sochat, V. & Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLoS One* **12**, e0177459 (2017).
37. Gaedigk, A. *et al*. Characterization of reference materials for genetic testing of CYP2D6 alleles: a GeT-RM collaborative project. *J. Mol. Diagn.* **21**, 1034–1052 (2019).
38. Twesigomwe, D., Wright, G.E.B., Drögemöller, B.I., da Rocha, J., Lombard, Z. & Hazelhurst, S. A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on *CYP2D6* genotyping. *NPJ Genom. Med.* **5**, 30 (2020).
39. Eggertsson, H.P. *et al*. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
40. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
41. Karczewski, K.J. *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
42. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
43. Collins, R.L. *et al*. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
44. EGFR gnomAD SVs v2.1 <https://gnomad.broadinstitute.org/gene/ENSG00000146648?dataset=gnomad_sv_r2_1> Accessed December 20, 2020.
45. VDR gnomAD SVs v2.1 <https://gnomad.broadinstitute.org/gene/ENSG00000111424?dataset=gnomad_sv_r2_1> Accessed December 20, 2020.
46. Pratt, V.M. *et al*. Characterization of 137 genomic DNA reference materials for 28 pharmacogenetic genes: a GeT-RM collaborative project. *J. Mol. Diagn.* **18**, 109–123 (2016).
47. Gordon, A.S., Fulton, R.S., Qin, X., Mardis, E.R., Nickerson, D.A. & Scherer, S. PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenet. Genomics* **26**, 161–168 (2016).

Appendix F: Plagiarism declaration and Turnitin report

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

Plagiarism Declaration

I David Twesigomwe (Student number: 1989491) am a student registered for the degree of Doctor of Philosopy 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" software indicating the level of plagiarism in my research document.

Signature: _____ Date: 01/06/2022

Turnitin report

1989491:David_Twesigomwe_Thesis_turnitin.pdf

ORIGINALITY REPORT

13%	11%	4%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS