

# Targeted ultra-deep sequencing of a South African Bantu-speaking cohort to comprehensively map and characterize common and novel variants in 65 pharmacologically-related genes

Sibongile Tshabalala<sup>a,b,e</sup>, Ananyo Choudhury<sup>b</sup>, Natasha Beeton-Kempen<sup>e</sup>, Neil Martinson<sup>c,d</sup>, Michèle Ramsay<sup>a,b</sup> and Dalu Mancama<sup>e</sup>

**Background** African populations are characterised by high genetic diversity, which provides opportunities for discovering and elucidating novel variants of clinical importance, especially those affecting therapeutic outcome. Significantly more knowledge is however needed before such populations can take full advantage of the advances in precision medicine. Coupled with the need to concisely map and better understand the pharmacological implications of genetic diversity in populations of sub-Saharan African ancestry, the aim of this study was to identify and characterize known and novel variants present within 65 important absorption, distribution, metabolism and excretion genes.

**Patients and methods** Targeted ultra-deep next-generation sequencing was used to screen a cohort of 40 South African individuals of Bantu ancestry.

**Results** We identified a total of 1662 variants of which 129 are novel. Moreover, out of the 1662 variants 22 represent potential loss-of-function variants. A high level of allele frequency differentiation was observed for variants identified in this study when compared with other populations. Notably, on the basis of prior studies, many appear to be pharmacologically important in the pharmacokinetics of a broad range of drugs, including

antiretrovirals, chemotherapeutic drugs, antiepileptics, antidepressants, and anticoagulants. An in-depth analysis was undertaken to interrogate the pharmacogenetic implications of this genetic diversity.

**Conclusion** Despite the new insights gained from this study, the work illustrates that a more comprehensive understanding of population-specific differences is needed to facilitate the development of pharmacogenetic-based interventions for optimal drug therapy in patients of African ancestry. *Pharmacogenetics and Genomics* 2019: 167–178. Copyright © 2019 The Author(s). Published by Wolters Kluwer Health, Inc.

*Pharmacogenetics and Genomics* 2019, 27:167–178

**Keywords:** Bantu-ancestry, genetic diversity and pharmacogenetics, next generation sequencing, novel variants

<sup>a</sup>Division of Human Genetics, National Health Laboratory Service, School of Pathology, Faculty of Health Sciences, <sup>b</sup>Sydney Brenner Institute for Molecular Bioscience (SBIMB), Faculty of Health Sciences, <sup>c</sup>Perinatal HIV Research Unit, Baragwanath Hospital and Faculty of Health Sciences, <sup>d</sup>Department of Internal Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg and <sup>e</sup>CSIR Biosciences Unit, Pretoria, South Africa.

Correspondence to Sibongile Tshabalala, MSc, CSIR Biosciences, PO Box 395, Pretoria 0001, South Africa  
Tel: +27 128 412 168; fax: +27 128 413 105; e-mail: stshabalala1@csir.co.za

Received 28 February 2019 Accepted 16 April 2019

## Introduction

There is increased interest in the genetic diversity of African populations and their contribution to understanding human evolution and disease susceptibility [1–5]. At the same time, it has become clear that genetic diversity significantly influences how populations across the continent respond to drug-based therapy [6–11]. The extent and impact of this influence remain unclear, necessitating studies for higher resolution mapping of

pharmacogenetic profiles in these populations. Genetic variants in pathways involved in drug absorption, distribution, metabolism and excretion (ADME) represent the main targets of pharmacogenetic studies, which seek to explain inter-patient treatment variability arising from a genetic predisposition. The vast majority of studies have been carried out on populations of European, Asian and African American ancestry, showing the presence of multiple variants that influence drug efficacy and safety across the treatment spectrum. Consequently, knowledge is biased towards facilitating pharmacogenetic-based interventions for non-African populations [12–14] and limited pharmacogenetic data are available for populations of African ancestry [6,15,16].

Populations of African ancestry harbour a large proportion of known variants with significant novel variant discovery

This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NCND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, [www.pharmacogeneticsandgenomics.com](http://www.pharmacogeneticsandgenomics.com).

in new sequencing studies [1,17,18], reflecting greater genetic diversity compared with other populations. Traditionally, African American, Yoruba, and Luhya populations have been used as representative populations for African ancestry [19,20]; however, recent studies have highlighted key differences in allele frequencies within sub-Saharan African (SSA) populations, suggesting that they are not ideal proxy populations of all Africans [1,21–23]. The differences reflect population admixture, random drift and/or selection, highlighting the need to study more African populations to identify population-specific variants that influence drug treatment.

Reflecting the continent's high HIV burden, African studies have traditionally focused on variants known to affect antiretroviral drug efficacy and outcome. Efavirenz (EFV) represents one such example, and the association of *CYP2B6* variants with the slow metabolism of EFV has been well described across the continent [9,24–27]. The prevalence of *CYP2B6*\*6 is significantly higher in African populations [23]. Other key findings in SSA have included association between *ABCB1* variants and differing EFV plasma concentrations [10], association between nevirapine-associated hepatotoxicity and human leukocyte antigen variants (*HLA*)-*DRB1*\*0102 and *HLA-B*\*5801 [28], and finally, the possible effect of *CYP3A4*, *CYP2B6* and *ABCB1* gene variants on antiretroviral drug efficacy through their influence on CD4<sup>+</sup> recovery rates [6].

The influence of variants on antimicrobial therapy has also received increased attention. Studies have reported associations between *SLCO1B1* variants and increased rifampicin clearance in South African patients [29], whereas *NAT2* variants show association with drug-induced liver injury (DILI) [30]. A *NAT2* genotype-guided regimen has thus been proposed to reduce isoniazid-associated DILI and early treatment failure [30]. Similarly, among Ethiopian patients, variants have been reported for the *NAT2* slow-acetylator phenotype and the *ABCB1* 3435TT genotype; these represent potential biomarkers for predisposition to DILI in tuberculosis (TB)-HIV coinfecting individuals [31]. Nonetheless, these dosing recommendations are still primarily based on clinical trials conducted in European or Asian populations, despite pharmacogenetic studies in African populations uncovering notable population differences in drug metabolism and efficacy [32].

South Africa hosts unique genetic diversity present in indigenous Khoesan and major Bantu-speaking populations [4]. The Bantu-speaking population is postulated to have originated from West Africa, from where it spread throughout SSA as supported by various genetic studies [33,34]. This diversity provides an opportunity to discover novel variants affecting the therapeutic outcome, and to begin developing more tailored pharmacogenetic

interventions. Given this, and driven by the need for more accurate mapping of genetic diversity in SSA populations and more specifically genes involved in xenobiotic metabolism, we used targeted next-generation sequencing to comprehensively screen the exons of 65 genes involved in the metabolism and therapeutic outcome of the majority of drugs in use today. This work represents, to our knowledge, the first study to comprehensively map at ultra-deep coverage the key genes of pharmacological relevance in individuals of Bantu ancestry. Furthermore, we explore and discuss the implications of these findings for future pharmacogenetic studies and strategies aimed at improving treatment outcome among patients of SSA origin.

## Patients and methods

### Study participants

The cohort comprised 40 unrelated Black South African individuals of Bantu ancestry, recruited from the Soweto catchment area through the Perinatal HIV Research Unit, Chris Hani Baragwanath Hospital. This number was sufficient to accurately map the frequency of common (frequency >0.01) variants and, as we show, also enabled the detection of relatively rare variants. The participants were HIV positive, but otherwise healthy patients (25 women and 15 men). The mean age of the participants was 38.6 years: 36.9 years for women and 41.4 years for men. Blood samples were collected with consent and ethics approval for the study was granted by the CSIR Ethics Committee (ref: 58/2013) and the University of Witwatersrand Human Research Ethics Committee (ref: 1201612). Genomic DNA was extracted from buffy coat samples using a Qiagen DNeasy blood and tissue kit (Qiagen, Hilden, Germany). Genomic DNA quality and quantity were analysed using a NanoDrop spectrophotometer (ThermoFisher Scientific, Massachusetts, USA) and the Qubit dsDNA HS Assay Kit using a Qubit 2.0 fluorometer (ThermoFisher Scientific).

### Next-generation sequencing

Targeted sequencing of exomes and immediately adjacent noncoding regions was performed on the 65 ADME genes listed in Table 1; primers were designed using Ion AmpliSeq Designer Pipeline version 4.0 [<http://www.ampliseq.com>] (ThermoFisher Scientific). Details of the Ion AmpliSeq panel are accessible through the link in Table 1. The primer panel covered 98.6% of the targeted exons. Sequencing was performed according to the manufacturer's recommendations (Supplementary Fig. 1, Supplemental digital content 1, <http://links.lww.com/FPC/B353>). Libraries were prepared at the CSIR and were sequenced at the National Genomics Infrastructure/SciLifeLab (Uppsala University, Sweden), on an Ion S5 sequencer using two Ion 530 chips (ThermoFisher Scientific).

**Table 1 Ion AmpliSeq gene panel<sup>a</sup> list summarizing the genes investigated in this study according to biological function**

Category or function	Genes
Transporter	<i>ABCC10, ABCB1, ABCC2, ABCC4, ABCG2, SLC15A2, SLC22A1, SLC22A11, SLC22A6, SLC22A8, SLC35D1, SLCO1A2, SLCO1B1, SLCO1B3, SLCO2A1, SLCO2B1, SLCO3A1, SLCO4A1, SLCO4C1</i>
Drug-metabolizing enzyme	<i>CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP2E1, CYP3A4, CYP3A5, CYP3A7, CYP8B1, DPYD, ANPEP, GGT1, GSTM1, UGT1A1, CES1, CES2, FMO1, FMO2, FMO3, NAT1, NAT2, SULT1A1</i>
Renal function	<i>B2M, CLU, LCN2, TIMP1, HAVCR1 (or KIM)</i>
Immune function	<i>HLAB, HLAC, HLADRB1, IL18</i>
Transcription factor	<i>AHR, NR1I2, NR1I3</i>
Other	<i>AGTR2, AGTRAP, ACE, FGA, FGG, TFF3, POR, NAGA</i>

<sup>a</sup>Complete details of the ion AmpliSeq gene panel are available through the following link: <https://www.ampliseq.com/protected/shared/viewSharedDesign.action?key=809e12a81cb5fe8caaa06e417c9f6c2cb500ea1966ad1332dfdde11c91f3c2b1>.

### Data processing and variant filtering

Raw data were processed in Torrent Suite Software version 5.0.2 (ThermoFisher Scientific, Massachusetts, USA) on the basis of default processing and quality filtering parameters. Sequences were aligned to the hg19 reference genome (assembly accession: GCF\_000001405.13 [35]) for mapping, base calling, and variant calling. Variant calling for homozygous or heterozygous single nucleotide variants was subsequently performed, together with the analysis of other changes such as insertions or deletions. Data were exported in the form of binary alignment map files and variant call format files. Golden Helix Genome Browser and SNP Variation Suite version 8.4.4 (SVS) (Golden Helix, Bozeman, Montana, USA) were used to mine, visualize and provide descriptive data statistics; variants that did not have a minimum genotype quality score of at least 15 and a read depth at least 10 were excluded from further analysis. Duplicates were removed and only annotation relevant to the 65 genes was retained. To enable a comparative analysis against existing data, Plink2.0 [36] was used to extract data for the SNPs from three datasets [the 1000 Genomes Project phase 3 (KGP) [20,37], the African Genome Variation Project (AGVP-ZUL) [1] and the Exome Aggregation Consortium (ExAC) 0.3 (African) [38]].

### Assessment of the potential functional impact of variants

For *in-silico* functional analysis, PolyPhen 2 (Polymorphism Phenotyping v2) and SIFT (Sorts Intolerant From Tolerant amino acid substitutions) algorithms, together with the ExAC Variant Effect Predictor Annotations 0.3 database, were used within SVS to identify and classify the potential functional impact of nonsynonymous and loss-of-function (LoF) variants. Additional annotation information was obtained from dbSNP using reference SNP cluster reports for known variants and cross-reference checks using Ensemble Variant Effect Predictor.

### Assessment of novel variants

To establish the presence of novel variants, variant filtering was performed by excluding variants contained within KGP, dbSNP common 147 (NCBI) and ExAC databases. The gene association of each novel variant was subsequently determined.

### Population structure

To determine the population structure of our cohort, 68 SNPs that fell below a call rate threshold of 0.8 at a coverage at least 30X were omitted. Principal component analysis (PCA) was carried out using Plink v2.0; plots were visualized with Genesis [39]. To identify SNPs and genomic regions showing strong population differentiation, we estimated the fixation index ( $F_{ST}$ ) for each individual SNP as well as across 10 kb genomic regions between the study population and AGVP-ZUL (South-African), KGP-YRI (West-African), KGP-CHB (Chinese) and KGP-CEU (European) populations using variant call format tools [40]. As some of the genomic regions were found to contain too few SNPs to provide meaningful averages, 10 kb genomic regions containing less than five SNPs were excluded from the analysis.

### Pharmacogenetic assessment of variants

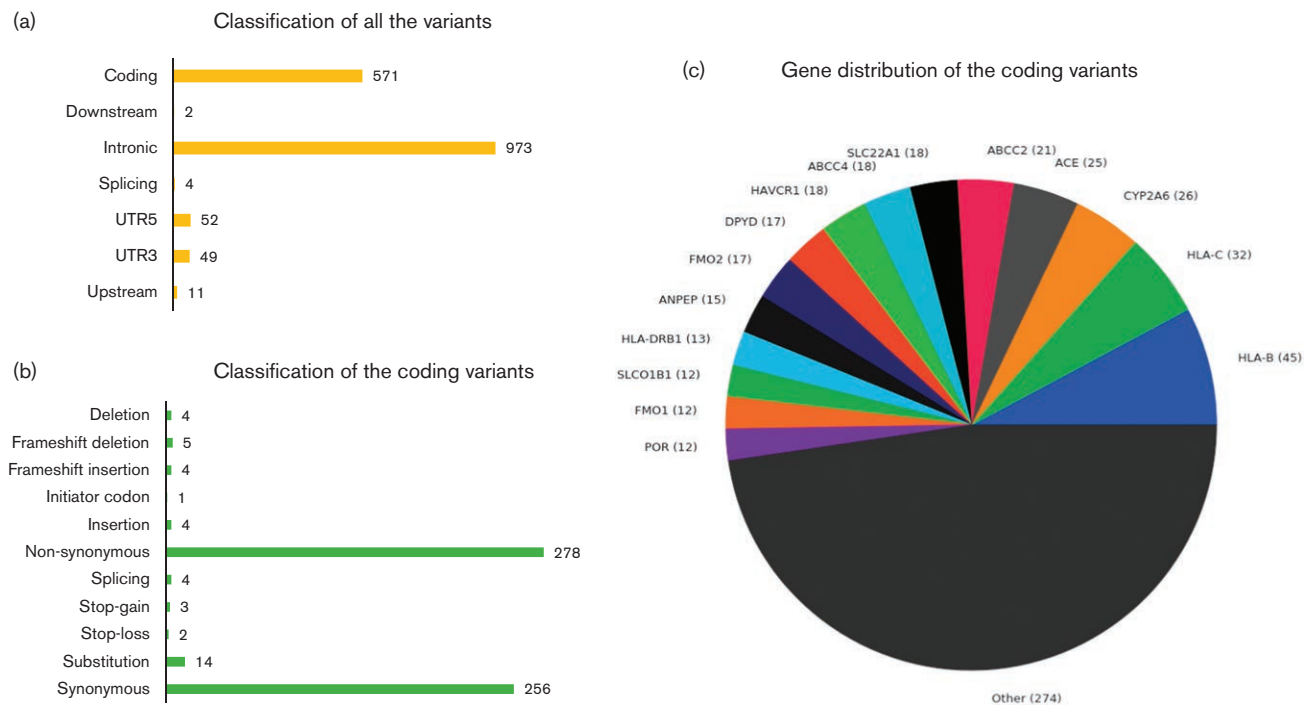
To delineate the potential impact that the variants confer on drug pharmacology, the Pharmacogenomics Knowledgebase (PharmGKB) database (<http://www.pharmgkb.org/view/vips.jsp>, accessed 12 July 2017) was used to identify those documented to significantly influence drug pharmacology. All variants were reviewed against the PharmGKB data; however, for brevity, priority was assigned to assessing in more detail variants whose clinical annotation level of evidence is moderate (levels 2A and 2B) to high (levels 1A and 1B). In the case of the novel variants, those predicted to result in significant functional change were evaluated for potential pharmacological relevance.

## Results

### Sequencing and variant discovery

The sequencing libraries generated an average of 17 068 182 reads per chip, representing 3.93 Gbp of sequence data across both chips. The mean read length was 237 bp. The average sequencing coverage was ~900× (Supplementary Table 1, Supplemental digital content 2, <http://links.lww.com/FPC/B354>). Data were analysed for high-quality variants in SVS by eliminating variants with a genotype quality score of at least 15 and a read depth less than 10. This resulted in the identification of 1662 high-quality variants (Fig. 1 and Supplementary Table 2, Supplemental digital content 3, <http://links.lww.com/FPC/>

Fig. 1



Variant classification. (a) Positional classification relative to the transcript of each variant ( $n=1662$ ). (b) Nature of the 575 variants affecting nucleotides within the coding and splicing region of each transcript. (c) Number of coding and splice variants within each respective gene. Genes with less than 12 variants each are grouped together as fanalysis (Fig. 1). Further 'other'. UTR, untranslated region.

B355) from 1996 variants originally identified. In addition, three variants more than 30 bp in size were observed, including a 32 bp deletion in *POR*, a 65 bp deletion in *HAVCR1* and a 51 bp insertion in *HAVCR1* (Supplementary Table 2, Supplemental digital content 3, <http://links.lww.com/FPC/B355>). Variant data are accessible through NCBI dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP>) by entering the submitter handle CSIRBIOHTS in the Search Entrez input box or through the full link supplied in Supplementary Table 2 (Supplemental digital content 3, <http://links.lww.com/FPC/B355>). Variants were distributed across the cohort at an average of 232 variants per individual, with the majority occurring in two or more individuals. High unique variant counts were observed in genes involved in drug transport including *ABCC4* and *ABCC2* (70 and 49 unique variants, respectively) and metabolism for example *ACE* (52 unique variants). Interestingly, the highest diversity was observed in genes often implicated in drug hypersensitivity (*HLA-B* and *HLA-C*, 86 and 85 unique variants, respectively). Conversely, highly conserved genes such as *IL18* and *DPYD* had fewer variants (two and three unique variants, respectively).

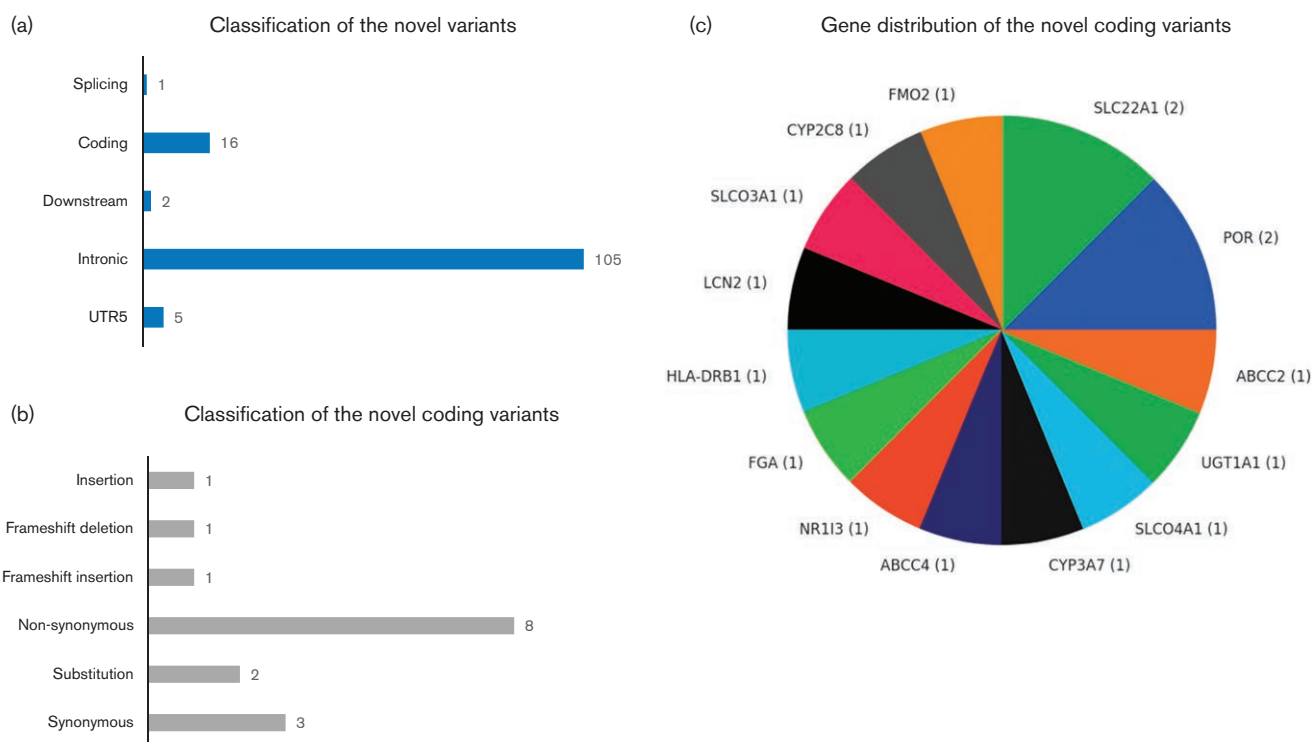
#### Variant identification, classification, and functional analysis

A high proportion of the total variability (973 variants, 58.5%) occurred in intronic sequences (Fig. 1 and

Supplementary Fig. 2, Supplemental digital content 1, <http://links.lww.com/FPC/B353>); the identification of intronic variants reflects the design of the Ion AmpliSeq panel, which achieves optimal exon coverage by extending target amplification to include flanking intronic sequences (refer to AmpliSeq panel design link, Table 1). Coding sequences were more conserved, harbouring 571 (34.4%) variants (Fig. 1 and Supplementary Table 2, Supplemental digital content 2, <http://links.lww.com/FPC/B355>). The remaining variants were largely localized to the 3'-untranslated region [49 (2.9%)], 5'-UTR [52 (3.1%)] or other [17 (1.0%)] regions. To identify the previously undescribed variants, those not in dbSNP were cross-referenced against those contained in the KGP, AGVP and ExAC databases. This resulted in the identification of 129 novel variants that were distributed across 49 ADME genes (Fig. 2 and Supplementary Table 3, Supplemental digital content 3, <http://links.lww.com/FPC/B356>).

In total, 256 nonsynonymous variants were identified through *in-silico* analysis (Fig. 1). Further analysis predicted 22 potential LoF variants representing either a frameshift, initiator codon, splice acceptor/donor or stop gained/lost change (Table 2). The clinical significance of the majority of these remains to be established; however, the importance of LoF variants such as *CYP2D6*\*4

Fig. 2



Novel variant classification. (a) Classification of each novel variant relative to their transcripts ( $n=129$ ). (b) Nature of the 16 novel coding variants affecting nucleotides within the coding region of each transcript. (c) Genes and the corresponding number of variants occurring within each respective gene. UTR, untranslated region.

(rs3892097) [minor allele frequency (MAF)=0.04 in our study population] is well established. This variant results in a poor metabolizer phenotype because of premature mRNA termination that is associated with poor tamoxifen metabolism during anticancer treatment and adverse drug reactions during the treatment of depression [41–44]. In addition to such variants, analysis of nonsynonymous variants identified 17 possibly damaging and/or deleterious consensus variants, and 41 that are probably damaging and/or deleterious (Table 3). For the majority of these, functional studies are required to determine the extent to which they contribute towards drug pharmacology.

### Population structure and differentiation

To study the cohort's population affinities and to detect possible population structure in the cohort, PCA was carried out on a set of 793 independent SNPs pruned down (at an  $r^2$  cutoff of 0.6) from the set of 968 SNPs that were found across all three datasets: KGP [17], AGVP [1] and our dataset. For this analysis, we included two populations from the AGVP (Zulu, and Wolayta from Ethiopia) and the KGP populations (YRI, LWK, ASW, CEU, TSI, CHB and JPT). The quality/resolution of PCA depends on the number of SNPs included in the

study. Although many PCAs in recent population genetic studies are based on hundreds of thousands of SNPs [1,45,46], we could find a clear separation between populations from Asia, Africa, and Europe, despite the use of a relatively small number of SNPs (Supplementary Fig. 2, Supplemental digital content 1, <http://links.lww.com/FPC/B353>). The Wolyata and ASW, as expected because of strong Eurasian admixture, clustered between African and European populations [1]. Our study cohort showed overlap with the other three Bantu-speaking groups (Zulu, LWK, and YRI) (Supplementary Fig. 2, Supplemental digital content 1, <http://links.lww.com/FPC/B353>). Moreover, the analysis suggested four of the individuals to have potential Eurasian ancestry possibly as a consequence of a relatively recent admixture, which is not uncommon in various populations from these regions [45,46] (Supplementary Fig. 2, Supplemental digital content 1, <http://links.lww.com/FPC/B353>). Another possible source of ancestry in the cohort (particularly for these four individuals) was Khoesan; however, potential Khoesan admixture could not be investigated because of little overlap between the SNPs sequenced in our study and those genotyped in Schlebusch *et al.* [46]. Genes with important biological functions, such as ADME genes, are highly conserved with a relatively low genetic variation.

**Table 2 Annotation of loss-of-function variants**

Gene symbol	Variant genome coordinates	dbSNP rs number	Reference allele	Alternate allele	Gene region	Consequence	Protein modification <sup>a</sup>	Bantu cohort frequency of alternate allele (global frequency) <sup>b</sup>
AGTRAP	1:11808597	rs771467011	C	T	Exon	Stop gained	p.Leu98=	0.0375 (0.0008)
AGTRAP	1:11810271	rs531367893	-	CCCTGCC	Exon	Frameshift variant	-	0.0125 (0.0048)
FMO2	1:171165803	rs283698680	G	-	Exon	Frameshift variant	p.Val113Ter	0.225 (0.1028)
FMO2	1:171173121	rs2020866	C	T	Exon	Stop gained	p.Arg249Ter	0.0125 (0.0046)
FMO2	1:171178090	rs6661174	T	C	Exon	Stop lost	p.Ter472GlnextTer64	0.2125 (0.0463)
FMO1	1:171254588	rs60639054	C	T	Exon	Stop gained	p.Arg502Ter	0.012 (0.0040)
SLC15A2	3:121613361	-	-	T	Exon	Frameshift variant	p.Ser15Phefs	0.0125 (NA)
HAVCRT	5:156456731	rs111759551	G	T	Exon	Stop gained	p.Cys358Ter	0.025 (0.0114)
HLA-C	6:31238001	-	-	C	Exon	Frameshift variant	p.Thr295Hisfs	0.025 (NA)
HLA-B	6:31323953	-	-	G	Exon	Frameshift variant	p.Glu204Alafs	0.075 (NA)
HLA-B	6:31324176	-	-	-	Exon	Frameshift variant	p.Pro129=	0.0125 (NA)
SLC22A1	6:160543329	rs111863691	C	G	Exon	Frameshift variant	p.Cys121Trpfs	0.0125 (NA)
SLC22A1	6:160560898	rs113569197	TGGTAAGT	-	Exon	Splice donor variant	-	0.3375 (0.8895)
CYP3A5	7:99250394	rs41303343	-	A	Exon	Frameshift variant	p.Thr346Tyrfs	0.0875 (0.0315)
CYP3A5	7:99269397	rs6977165	T	C	Exon	Stop lost	-	0.1 (0.0649)
CYP2C8	10:96827120	rs11572078	-	A	Exon	Frameshift variant	-	0.2 (0.3498)
CYP2C8	10:96829159	rs142470035	T	C	Exon	Start lost	p.Met1?	0.0125 (0.0012)
ABCC2	10:101577129	-	TT	-	Intron	Splice donor variant	p.Phe721Trpfs	0.0125 (NA)
CES1	16:55866915	rs139063675	C	T	Exon	Frameshift variant	-	0.0375 (0.0008)
GGT1	22:25023506	-	G	-	Exon	Frameshift variant	p.Asp377Metfs	0.0125 (NA)
NAGA	22:42461731	-	CATCTGGTA	-	Intron	Splice donor variant	-	0.0125 (NA)
CYP2D6	22:42524947	rs3892097	C	T	Intron	Splice acceptor variant	-	0.0375 (0.0931)

<sup>a</sup>Key: fs, frameshift; Ter, termination.

<sup>b</sup>Global alternate allele frequencies are based on data from the 1000 Genomes Project phase 3.

**Table 3 Functional analysis of variants identified in the cohort**

Variant prediction tool	Prediction	Number of variants
PolyPhen 2.0	Possibly damaging (0.8–0.9)	27
	Probably damaging ( $\geq 0.9$ )	46
SIFT	Deleterious ( $\leq 0.05$ )	97
Consensus of PolyPhen and SIFT	Possibly damaging and deleterious	17
	Probably damaging and deleterious	41

This was observed to be the case for the more closely related Bantu-speaking populations (Zulu, LWK and YRI), resulting in minimal separation as determined by PCA. Analysis of a larger subset of more diverse genes would thus be expected to identify distinguishing signatures of each population among these populations.

For the SNPs that were shared between our cohort and other datasets, we estimated the  $F_{ST}$  for each in comparison with populations from the same geographic area [AGVP-ZUL and Central-West Africa (KGP-YRI)], and with other continents (KGP-CEU and KGP-CHB). Supplementary Fig. 3 (Supplemental digital content 1, <http://links.lww.com/FPC/B353>), illustrates the number of SNPs with  $F_{ST}$  scores higher than 0.15 (moderate genetic differentiation) and 0.25 (high genetic differentiation) in the various population comparisons. As expected, few SNPs were found to show very high  $F_{ST}$  scores between the cohort and other African populations (Supplementary Fig. 3, Supplemental digital content 1, <http://links.lww.com/FPC/B353>). A detailed list of SNPs is provided in Supplementary Table 4 (Supplemental digital content 5, <http://links.lww.com/FPC/B357>). A similar analysis of weighted  $F_{ST}$  across 10kb blocks along the genome identified regions in chromosomes 1 (*DPYD*), 6 (*HLA-C*) and 13 (*ABCC4*) with high differentiation between our population and other African populations (Fig. 3 and Supplementary Table 5, Supplemental digital content 6, <http://links.lww.com/FPC/B358>).

### Pharmacological impact of the variants

Gene variants were examined for their potential pharmacological impact using data contained within the PharmGKB database. Allele frequency data representing 34 key variants, located within 16 important pharmacogenes, are summarized in Fig. 4. Assessment of these variants was based primarily on their relevance to drugs commonly prescribed within the broader population of our representative cohort, including notably those used to treat HIV and TB. Among these variants, one of the most frequent variants [rs1208 (*NAT2\*12*)] occurred at a MAF of 0.389, being comparatively common in ethnically related cohorts but relatively rare in others, for example, within the Chinese population (Fig. 4). The variant is associated with a rapid acetylator phenotype, affecting drugs such as isoniazid, sulphamethazine, sulphamethoxazole and trimethoprim commonly used to

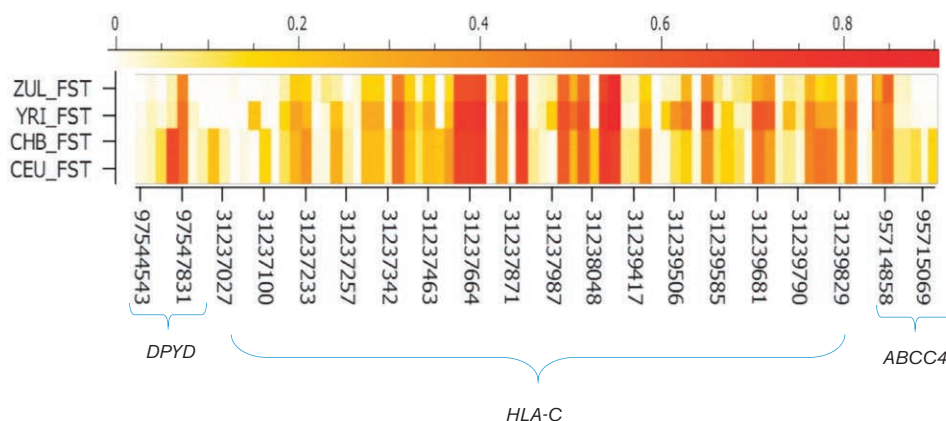
treat bacterial infections including TB. In contrast, a well-documented *ABCB1* variant (rs2032582, level 2A) is relatively rare among African populations (including our cohort), but frequent in individuals of Chinese (MAF=0.493) and European (MAF=0.433) descent. Close to two-thirds [22/34 (64.7%)] of the variants were relatively common (MAF $\geq 0.10$ ) in our cohort. Of these, *CYP2B6\*6* (rs3745274, MAF=0.264) is implicated widely in dosage-based variant–drug interactions (level 1B), notably during treatment of HIV infection with EFV. Several of the less frequent variants identified in our study have known or predicted pharmacogenomic relevance and some are either incorporated into prescribing guidelines or health systems guidelines [level 1A, namely *CYP2C19\*2* (rs4244285), *CYP2D6\*4* (rs3892097) and *SLCO1B1\*5* (rs4149056)], highlighting their potential clinical impact. For example, a novel, moderately frequent (MAF=0.075) missense *SLCO3A1* variant (15:92694224 T/C) was assigned a deleterious (SIFT score 0)/probably damaging (PolyPhen 2.0 score 0.99) phenotype, and a rarer (MAF=0.013), novel missense variant (2:234669414 G/T) in *UGT1A8* had a PolyPhen 2.0 score of 1. The functional validation of these variants remains to be performed.

### Discussion

Using high-throughput targeted sequencing, we sequenced 65 key ADME-related genes in 40 Bantu ancestry individuals and identified 1662 high-confidence unique variants, of which 129 were novel.

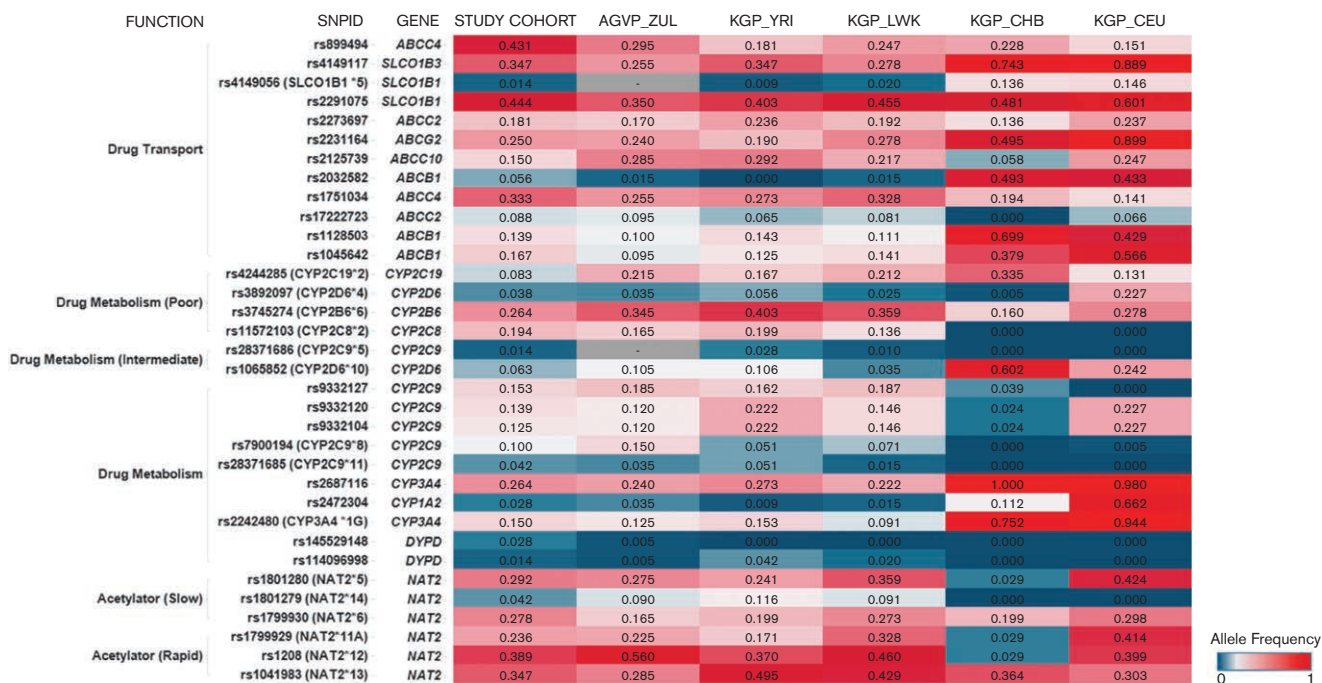
PCA analysis demonstrated that our cohort was most closely related to the Bantu-speaker populations represented in the AGVP and KGP (Supplementary Fig. 2, Supplemental digital content 1, <http://links.lww.com/FPC/B353>). Through  $F_{ST}$  analysis of 10kb genomic regions, we investigated the extent of genetic differentiation with other populations, and several putative signatures of positive selection were identified among the pharmacogenes. Clear distinctions were observed between African populations and those of European and Asian descent. A broad range of variants within these clusters are known to significantly influence drug therapy; this genetic heterogeneity may explain the differences in the response among populations to the same drug, supporting the hypothesis that genetic heterogeneity may underlie notable discrepancies where these populations respond differently to the same drugs. Interestingly, evidence of selection for *HLA* variants was observed among African populations, possibly reflecting benefits to immune function to be gained from such diversity, given the nature of the continent's diverse disease burden. Conversely,  $F_{ST}$  analysis identified relatively high differentiation between our population and other African populations for gene regions along chromosomes 1 (*DPYD*), chromosome 6 (*HLA-C*) and 13 (*ABCC4*). These genes play important roles in pyrimidine metabolism, HIV

Fig. 3



Population differentiation. Genomic regions showing strong population differentiation were identified by estimating the weighted mean fixation index ( $F_{ST}$ ), across 10 kb nonoverlapping windows, between the study population and the AGVP-ZUL (South-African), KGP-YRI (West-African), KGP-CHB (Chinese) and KGP-CEU (European) populations.  $F_{ST}$  values for SNPs from three gene regions showing maximum differentiation are shown in the heat map. The bar at the top of the figure shows the  $F_{ST}$  scale. The genomic coordinates of every fourth/fifth SNP are shown. Variations in *DPYD* are implicated in the efficacy of fluorouracil (anticancer treatment) [42], variations in *ABCC4* are implicated in tenofovir (antiretroviral treatment) efficacy and variations in *HLA-C* are implicated in ARV, anti-TB and anti-inflammation therapy. ARV, antiretrovirals; HLA, human leukocyte antigen; SNP, single nucleotide polymorphism; TB, tuberculosis..

Fig. 4



A comparison of variant allele frequencies with potential clinical significance across multiple populations. The heat map illustrates the allele frequency of 34 clinically relevant variants as ascribed within PharmGKB. Allele frequencies for the cohort are compared with those from three other African populations (AGVP-Zul, KGP-YRI and KGP-LWK) and a European (CEU) population. Frequencies are depicted across a shade spectrum ranging from relatively rare (blue) to relatively frequent (red).

disease progression and organic anion transport, respectively, and variants thereof are implicated in adverse drug reactions associated with antiretroviral therapy and oncotherapy [47–49].

While investigating the evidence for population-based ADME variation, we confirmed the presence of variants of clinical importance, furthermore observing notable allele frequency differences in our cohort compared



with other populations. This was observed for multiple variants implicated in the pharmacology of antiretrovirals, antimicrobials, antimalarials, anticoagulants, chemotherapeutic drugs and antiepileptics. This is the case for variants in *CYP2B6* that are implicated in the altered metabolism of several drugs, notably EFV. Mirroring findings for several other African populations [23], and those of European (CEU) descent, *CYP2B6\*6* (rs3745274) was relatively more common (MAF=0.264) compared with the Han Chinese (CHB) populations (MAFs=0.004). A similar MAF (0.20) was noted in the Xhosa and Cape mixed-ancestry populations (CMA) [50]. This variant has increasingly been studied in Southern African populations, given the widespread use of EFV and is associated with susceptibility to EFV-induced adverse events [24,26,51]. In the case of another widely prescribed antiretroviral drug, tenofovir disoproxil fumarate (TDF), acute kidney injury (AKI) poses a notable challenge to HIV management [52–57]. The most common variant implicated in AKI, rs717620, located in the *ABCC2* gene [58–60], was rare in our cohort (MAF=0.013), suggesting that in the broader population, its influence may not be as widespread compared with other populations. In contrast, another commonly implicated *ABCC2* variant, rs2273697, showed moderate frequency (MAF=0.181). Other variants that affect TDF treatment were found to be more frequent in our cohort, such as rs1751034 located in *ABCC4* (MAF=0.333), which is associated with the increased intracellular concentration of TDF [48]. There is thus merit in prioritizing the analysis of such variants in the broader population to establish their clinical relevance to such therapy.

HIV management is increasingly complicated by concomitant diseases, in particular, TB co-infection. Coadministration of antiretrovirals and anti-TB drugs is associated with major complications, including immune reconstitution inflammatory response [61] and DILI. In South Africa, up to 8.3% of admissions and 2.9% of hospital deaths have been attributed to adverse reactions associated with TDF, rifampicin and co-trimoxazole coadministration [52,54,62]. Identification of valid genetic biomarkers that can guide treatment and prevent such outcomes has therefore become highly warranted. Isoniazid is commonly used as part of the anti-TB treatment regimen; variants of the *NAT2* gene are implicated widely in the variable metabolism rates observed for this drug because of rapid, intermediate or slow metabolism (acetylation) phenotypes [63]. Functional variants have been described for *NAT2*, with studies indicating significant differences in functional classes among ethnically diverse populations [63]. Rapid acetylators are individuals at risk of potential drug resistance [64] and slow acetylators show reduced drug clearance coupled with increased isoniazid and hydrazine exposure, and as a result are at increased risk of hepatotoxicity, liver injury or hepatitis induced by anti-TB treatment [63]. The incidence of

drug-induced hepatotoxicity in Africa ranges from 8 to 21.2% [65,66] in patients receiving anti-TB treatment. Interestingly, slow-acetylator phenotypes are most prevalent within the European and African populations [67]; variants representing this phenotype were comparatively common in our cohort (Fig. 4). Another relatively common important *NAT2* variant in our cohort (*NAT2\*11 A*; rs1799929) is associated with anti-TB treatment outcome. Recent data have shown how a *NAT2* genotype-guided regimen that includes this variant, combined with variants represented in *ABCB1* such as rs1045642, can reduce isoniazid-induced liver injury (DILI) and early treatment failure in TB-HIV coinfecting patients [30,31]. Given the increasingly widespread use of anti-TB drugs among Bantu-speaking patients, *NAT2* variants present important candidates for further studies to determine their clinical effect on TB drug outcome, even more so during concomitant antiretroviral use.

Effective malaria treatment represents another clinical intervention on the continent where optimal drug regimens are critical for success, given the rapid pace of infection and the potential for disease resistance following inadequate treatment. Variants in several genes encoding members of the cytochrome P450 (*CYP*) enzyme family are implicated widely in altered antimalarial metabolism, resulting for example in increased drug serum concentration levels because of poor metabolism [68,69]. This is the case for *CYP2C19*, where variable activity influences prescription guidelines for a number of drugs [70–72]. On the basis of *CYP2C19* activity levels, individuals can be classified as ultrarapid metabolizer, rapid metabolizer, extensive metabolizer, intermediate metabolizer or PM. Kaneko *et al.* [73] first identified an association between *CYP2C19* variants and the PM of proguanil. PM individuals have two LoF alleles (*\*2/\*2*, *\*2/\*3*, *\*3/\*3*), resulting in markedly reduced or absent *CYP2C19* activity. Conversely, UM individuals have two gain-of-function alleles (*\*1/\*17*, *\*17/\*17*), resulting in increased enzyme activity [70,74]. Of these, rs4244285 (*\*2*) occurs at a MAF of 0.083 in our cohort (Fig. 4). Another important allele implicated in the PM of antimalarials is *CYP2C8\*2* [69]. Several African studies have noted distinct inter-ethnic differences in the frequency of *CYP2C8\*2* [75,76]. We noted similar allele frequencies to a Bantu cohort from Botswana (0.194 and 0.175, respectively) [75]. Given the scale of antimalarial use in the region, these prevalence rates merit studies to assess the relevance of considering genotype status before treatment or prophylaxis. Such strategies have become increasingly important as efforts grow to eliminate the disease, where failed treatment because of suboptimal dosing could precipitate the emergence of resistant parasite strains [77].

Although infectious disease represents a significant proportion of the continent's disease burden, noncommunicable illnesses are increasingly important. An important

enzyme in this respect is CYP2C9, involved in the oxidation of drugs including warfarin, losartan, and phenytoin. *CYP2C9* variants, notably *CYP2C9\*2* (rs1799853), *CYP2C9\*3* (rs1057910), *CYP2C9\*5* (rs28371686), *CYP2C9\*8* (rs7900194) and *CYP2C9\*11* (rs28371685), are commonly associated with reduced warfarin clearance, affecting the dosing of this drug [78–81]. *CYP2C9\*2* and *CYP2C9\*3* are more frequent in European and American (including African-American) populations; however, these variants are rare in African populations (including our study population). In contrast, *CYP2C9\*5*, *CYP2C9\*8*, and *CYP2C9\*11* are more frequent in African populations (Fig. 4). Given their prevalence, such genetic biomarkers could be exploited for improving warfarin efficacy in populations of African ancestry [79,81].

Other important variants that warrant consideration for this population on the basis of their PharmGKB classification include rs4244285 (*CYP2C19\*2*, level 1A), which influences clopidogrel (cardiovascular disease) and amitriptyline (depression) efficacy, and rs1045642 (*ABCB1*, level 2A), which is associated with toxicity and adverse events during the treatment of lymphomas with methotrexate. Identification of the vast majority of important pharmacogene variants in individuals of Bantu ancestry now provides a broad basis for prioritizing the future investigation of these and other variants with a potential influence on drug treatment outcome for noncommunicable diseases in this population.

An important aspect of the study was the high number (129; 7.8%) of novel variants identified, raising the prospect of new phenotypes that significantly influence treatment outcome. Moreover, out of the 1662 variants, 22 were predicted to be LoF and thereby may impact drug metabolism and therapeutic response. Examples included a common frameshift variant (MAF = 0.23) and stop-loss variant (MAF = 0.21) in the flavin-containing monooxygenase 2 gene, which bioactivates substrates such as thioureas to sulphenic or sulphinic acid metabolites [82]. Similarly, less prevalent but equally important LoF variants were uncovered within genes encoding *CYP2D6*, *CYP3A5*, *CYP2C8* and *CES1* (Table 2). Although the predominant focus of pharmacogenetic studies has been to establish the influence of relatively common variants on treatment outcome, it is increasingly clear that relatively rare variants that confer significant functional change are also key to achieving the goals of precision medicine, particularly with respect to explaining the occurrence of less frequent clinical observations linked to drug use [12–14]. Thus, determining the biological effects of these changes will provide a clear understanding of their potential pharmacological roles, and the relevance of these changes towards improving drug-based therapy.

Despite the important new insights gained in this study through AmpliSeq-based sequencing, we acknowledge

the technology's limitations. One challenge is in completely capturing homopolymer sequences, that fortuitously are rare across the gene regions that we targeted. In addition, although it is expected that most pharmacogenetically important variants exist within coding regions, we would have missed potentially relevant variants in the intronic, upstream and regulatory regions. One example is *CYP2C19\*17* (rs12248560), an upstream variant associated with an UM phenotype. Similarly, copy number variants were not investigated. Although this study investigated the majority of the key ADME-related variants, future studies would benefit from the inclusion of variants in other similar genes that may be important to drug pharmacology. Notwithstanding this, we confirmed the presence of high pharmacogenetic diversity in an African population and highlighted the need for further research upon which to develop improved strategies for tailored pharmacological intervention.

### Conclusion

Populations across SSA are genetically diverse, but relatively little is known in terms of the extent to which inter-ethnic differences impact upon drug-based therapeutic outcome. We mapped the variant composition of 65 pharmacologically important genes in a cohort of Bantu ancestry, resulting in the identification of 1662 variants of high confidence, of which 129 were found to be novel. On the basis of *in-silico* analysis, several of these are predicted to result in functional changes, providing motivation for follow-up studies to characterize and determine their clinical pharmacological effects. Ultimately, validation of their clinical relevance or otherwise, in conjunction with our knowledge of the prevalence of known variants of clinical relevance, will prove instrumental in guiding new policies for drug selection and dosing in African populations on the basis of pharmacogenetic principles and strategies aimed at improving drug safety and efficacy.

### Acknowledgements

The authors thank Turflos Netshilindi for extracting the DNA, and Inger Jonasson, Susana Haggqvist and Adam Ameer at the National Genomics Infrastructure, SciLifeLab Department of Immunology, Genetics and Pathology, Uppsala University, Sweden, for sequencing our libraries on the Ion S5 sequencer. They also thank the nursing staff for patient recruitment and sample collection. They are grateful and indebted to all participants in this study.

Funding was provided by the Department of Science and Technology (grant # V6YET50) and a CSIR parliamentary (grant # V1YBT96) (N.B.K., D.M., S.T.). N.M. was supported by Perinatal HIV Research Unit funding. A.C. was supported by the AWI-Gen Collaborative Centre funded by the NIH (U54HG006938) as part of the

H3Africa Consortium. M.R. is a South African Research Chair in Genomics and Bioinformatics of African populations hosted by the University of the Witwatersrand, funded by the Department of Science and Technology and administered by the National Research Foundation of South Africa (NRF).

### Conflicts of interest

There are no conflicts of interest.

### References

- Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* 2015; **517**:327–332.
- May A, Hazelhurst S, Li Y, Norris SA, Govind N, Tikly M, *et al.* Genetic diversity in black South Africans from Soweto. *BMC Genomics* 2013; **14**:1.
- Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, *et al.* Admixture into and within sub-Saharan Africa. *eLife* 2016; **5**:e15266.
- Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* 2010; **463**:943–947.
- Ramsay M, Tiemessen CT, Choudhury A, Soodyall H. Africa: the next frontier for human disease gene discovery? *Hum Mol Genet* 2011; **20**:R214–R220.
- Parathyras J, Gebhardt S, Hillermann-Rebello R, Grobbelaar N, Venter M, Warnich L. A pharmacogenetic study of CD4 recovery in response to HIV antiretroviral therapy in two South African population groups. *J Hum Genet* 2009; **54**:261–265.
- Gorny M, Rohm S, Laer S, Morali N, Niehues T. Pharmacogenomic adaptation of antiretroviral therapy: overcoming the failure of lopinavir in an African infant with CYP2D6 ultrarapid metabolism. *Eur J Clin Pharmacol* 2010; **66**:107–108.
- Mukonzo JK, Owen JS, Ogwal-Okeng J, Kuteesa RB, Nanzigu S, Sewankambo N, *et al.* Pharmacogenetic-based efavirenz dose modification: suggestions for an African population and the different CYP2B6 genotypes. *PLoS One* 2014; **9**:e86919.
- Ngaimisi E, Habtewold A, Minzi O, Makonnen E, Mugusi S, Amogne W, *et al.* Importance of ethnicity, CYP2B6 and ABCB1 genotype for efavirenz pharmacokinetics and treatment outcomes: a parallel-group prospective cohort study in two sub-Saharan Africa populations. *PLoS One* 2013; **8**:e67946.
- Swart M, Ren Y, Smith P, Dandara C. ABCB1 4036A>G and 1236C>T polymorphisms affect plasma Efavirenz levels in South African HIV/AIDS patients. *Front Genet* 2012; **3**:236.
- Li J, Lao X, Zhang C, Tian L, Lu D, Xu S. Increased genetic diversity of ADME genes in African Americans compared with their putative ancestral source populations and implications for Pharmacogenomics. *BMC Genet* 2014; **15**:1.
- Fujikura K, Ingelman-Sundberg M, Lauschke VM. Genetic variation in the human cytochrome P450 supergene family. *Pharmacogenet Genomics* 2015; **25**:584–594.
- Kozyra M, Ingelman-Sundberg M, Lauschke VM. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. *Genet Med* 2017; **19**:20–29.
- Lauschke VM, Ingelman-Sundberg M. Precision medicine and rare genetic variants. *Trends Pharmacol Sci* 2016; **37**:85–86.
- Pillai G, Davies G, Denti P, Steimer JL, McIlerron H, Zvada S, *et al.* Pharmacometrics: opportunity for reducing disease burden in the developing world: the case of Africa. *CPT Pharmacometrics Syst Pharmacol* 2013; **2**:e69.
- Warnich L, Drogemoller BI, Pepper MS, Dandara C, Wright GE. Pharmacogenomic research in south africa: lessons learned and future opportunities in the rainbow nation. *Curr Pharmacogenomics Person Med* 2011; **9**:191–207.
- Auton A, Abecasis GR, Altshuler D, Durbin R, Bentley D, Chakravarti A, *et al.* A global reference for human genetic variation. *Nature* 2015; **526**:68–74.
- 1000 Genome Project Consortium, Abecasis GA, Auton D, Brooks A, Durbin LD, Gibbs RM, Hurler RA, McVean GA. A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**:1061–1073.
- Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**:52–58.
- Auton AB, Durbin LD, Garrison RM, Kang EP, Korbel HM, Marchini JO, *et al.* 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015; **526**:68–74.
- Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet* 2009; **41**:657–665.
- Li J, Menard V, Benish RL, Jurevic RJ, Guillemette C, Stoneking M, *et al.* Worldwide variation in human drug-metabolism enzyme genes CYP2B6 and UGT2B7: implications for HIV/AIDS treatment. *Pharmacogenomics* 2012; **13**:555–570.
- Aminkeng F, Ross CJD, Rassekh SR, Brunham LR, Sistonen J, Dube MP, *et al.* Higher frequency of genetic variants conferring increased risk for ADRs for commonly used drugs treating cancer, AIDS and tuberculosis in persons of African descent. *Pharmacogenomics J* 2013; **14**:160–170.
- Gounden V, van Niekerk C, Snyman T, George JA. Presence of the CYP2B6 516G>T polymorphism, increased plasma efavirenz concentrations and early neuropsychiatric side effects in South African HIV-infected patients. *AIDS Res Ther* 2010; **7**:32.
- Masebe TM, Bessong PO, Nwobegahay J, Ndip RN, Meyer D. Prevalence of MDR1 C3435T and CYP2B6 G516T polymorphisms among HIV-1 infected South African patients. *Dis Markers* 2012; **32**:43–50.
- Swart M, Skelton M, Ren Y, Smith P, Takuva S, Dandara C. High predictive value of CYP2B6 SNPs for steady-state plasma efavirenz levels in South African HIV/AIDS patients. *Pharmacogenet Genomics* 2013; **23**:415–427.
- Wang J, Sönnnerborg A, Rane A, Josephson F, Lundgren S, Ståhle L, *et al.* Identification of a novel specific CYP2B6 allele in Africans causing impaired metabolism of the HIV drug efavirenz. *Pharmacogenet Genomics* 2006; **16**:191–198.
- Phillips E, Bartlett JA, Sanne I, Lederman MM, Hinkle J, Rousseau F, *et al.* Associations between HLA-DRB1\*0102, HLA-B\*5801, and hepatotoxicity during initiation of nevirapine-containing regimens in South Africa. *J Acquir Immune Defic Syndr* 2013; **62**:e55-7.
- Chigutsa E, Visser ME, Swart EC, Denti P, Pushpakom S, Egan D, *et al.* The SLCO1B1 rs4149032 polymorphism is highly prevalent in South Africans and is associated with reduced rifampin concentrations: dosing implications. *Antimicrob Agents Chemother* 2011; **55**:4122–4127.
- Azuma J, Ohno M, Kubota R, Yokota S, Nagai T, Tsuyuguchi K, *et al.* NAT2 genotype guided regimen reduces isoniazid-induced liver injury and early treatment failure in the 6-month four-drug standard treatment of tuberculosis: a randomized controlled trial for pharmacogenetics-based therapy. *Eur J Clin Pharmacol* 2013; **69**:1091–1101.
- Yimer G, Ueda N, Habtewold A, Amogne W, Suda A, Riedel KD, *et al.* Pharmacogenetic & pharmacokinetic biomarker for efavirenz based ARV and rifampicin based anti-TB drug induced liver injury in TB-HIV infected patients. *PLoS One* 2011; **6**:e27810.
- Masimirembwa C, Hasler JA. Pharmacogenetics in Africa, an Opportunity for Appropriate Drug Dosage Regimens: on the Road to Personalized Healthcare. *CPT Pharmacometrics Syst Pharmacol* 2013; **2**:e45.
- Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, *et al.* Genetic and demographic implications of the bantu expansion: insights from human paternal lineages. *Mol Biol Evol* 2009; **26**:1581–1589.
- Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 2017; **356**:543–546.
- Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, *et al.* The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* 2009; **37**:D755–D761.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**:7.
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* 2015; **526**:75–81.
- Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, *et al.* Analysis of protein-coding genetic variation in 60,706 humans *Nature* 2015; **536**:285–291.
- B Robert, Scott H. Genesis 0.2.1 manual. 2015. Available at: <http://www.bioinf.wits.ac.za/software/genesis/> [Accessed 18 October 2017].
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* The variant call format and VCF tools. *Bioinformatics* 2011; **27**:2156–2158.

- 41 Schroth W, Goetz MP, Hamann U, Fasching PA, Schmidt M, Winter S, *et al.* Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. *JAMA* 2009; **302**:1429–1436.
- 42 Goetz MP, Rae JM, Suman VJ, Safgren SL, Ames MM, Visscher DW, *et al.* Pharmacogenetics of tamoxifen biotransformation is associated with clinical outcomes of efficacy and hot flashes. *J Clin Oncol* 2005; **23**:9312–9318.
- 43 Goetz MP, Sangkuhl K, Guchelaar HJ, Schwab M, Province M, Whirl-Carrillo M, *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and Tamoxifen Therapy. *Clin Pharmacol Ther* 2018; **103**:770–777.
- 44 Bijl MJ, Visser LE, Hofman A, Vulto AG, van Gelder T, Stricker BH, *et al.* Influence of the CYP2D6\*4 polymorphism on dose, switching and discontinuation of antidepressants. *Br J Clin Pharmacol* 2008; **65**:558–564.
- 45 Chimusa E, Meintjes A, T M, Mulder N, Seoighe C, Soodyall H, *et al.* A genomic portrait of haplotype diversity and signatures of selection in indigenous Southern African Populations. *PLoS Genet* 2015; **11**:e1005363.
- 46 Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, *et al.* Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science* 2012; **338**:374–379.
- 47 Kulpa DA, Collins KL. The emerging role of HLA-C in HIV-1 infection. *Immunology* 2011; **134**:116–122.
- 48 Kiser JJ, Aquilante CL, Anderson PL, King TM, Carten ML, Fletcher CV. Clinical and genetic determinants of intracellular tenofovir diphosphate concentrations in HIV-infected patients. *J Acquir Immune Defic Syndr* 2008; **47**:298–303.
- 49 Caudle KE, Thorn CF, Klein TE, Swen JJ, McLeod HL, Diasio RB, *et al.* Clinical Pharmacogenetics Implementation Consortium guidelines for dihydropyrimidine dehydrogenase genotype and fluoropyrimidine dosing. *Clin Pharmacol Ther* 2013; **94**:640–645.
- 50 Ikediobi O, Aouizerat B, Xiao Y, Gandhi M, Gebhardt S, Warnich L. Analysis of pharmacogenetic traits in two distinct South African populations. *Hum Genomics* 2011; **5**:265–282.
- 51 Nyakutira C, Roshammar D, Chigutsa E, Chonzi P, Ashton M, Nhachi C, *et al.* High prevalence of the CYP2B6 516G->T(\*6) variant and effect on the population pharmacokinetics of efavirenz in HIV/AIDS outpatients in Zimbabwe. *Eur J Clin Pharmacol* 2008; **64**:357–365.
- 52 Mouton JP, Njuguna C, Kramer N, Stewart A, Mehta U, Blockman M, *et al.* Adverse Drug Reactions Causing Admission to Medical Wards: A Cross-Sectional Survey at 4 Hospitals in South Africa. *Medicine* 2016; **95**:e3437.
- 53 Njuguna C, Stewart A, Mouton JP, Blockman M, Maartens G, Swart A, *et al.* Adverse drug reactions reported to a national HIV & tuberculosis health. *Drug Saf* 2016; **39**:159–169.
- 54 Mouton JP, Mehta U, Parrish AG, Wilson DPK, Stewart A, Njuguna CW, *et al.* Mortality from adverse drug reactions in adult medical inpatients at four hospitals in South Africa: a cross-sectional survey. *Br J Clin Pharmacol* 2015; **80**:818–826.
- 55 Seedat F, Martinson NA, Motlhaoleng K, Abraham P, Mancama D, Naicker S, *et al.* Acute kidney injury, risk factors and prognosis in hospitalized HIV-infected adults in South Africa, compared by tenofovir exposure. *AIDS Res Hum Retroviruses* 2016; **33**:33–40.
- 56 Waheed S, Attia D, Estrella MM, Zafar Y, Atta MG, Lucas GM, *et al.* Proximal tubular dysfunction and kidney injury associated with tenofovir in HIV patients: a case series. *Clin Kidney J* 2015; **8**:420–425.
- 57 Baxi SM, Scherzer R, Greenblatt RM, Minkoff H, Sharma A, Cohen M, *et al.* Higher tenofovir exposure is associated with longitudinal declines in kidney function in women living with HIV. *AIDS* 2016; **30**:609–618.
- 58 Rodríguez-Nóvoa S, Labarga P, Soriano V, Egan D, Albalater M, Morello J, *et al.* Predictors of kidney tubular dysfunction in HIV-infected patients treated with tenofovir: a pharmacogenetic study. *Clin Infect Dis* 2009; **48**:e108–e116.
- 59 Izzedine H, Hulot JS, Villard E, Goyenvalle C, Dominguez S, Ghosn J, *et al.* Association between ABCC2 gene haplotypes and tenofovir-induced proximal tubulopathy. *J Infect Dis* 2006; **194**:1481–1491.
- 60 Nishijima T, Komatsu H, Higasa K, Takano M, Tsuchiya K, Hayashida T, *et al.* Single nucleotide polymorphisms in ABCC2 associate with tenofovir-induced kidney tubular dysfunction in Japanese patients with HIV-1 infection: a pharmacogenetic study. *Clin Infect Dis* 2012; **55**:1558–1567.
- 61 McIlleron H, Meintjes G, Burman WJ, Maartens G. Complications of antiretroviral therapy in patients with tuberculosis: drug interactions, toxicity, and immune reconstitution inflammatory syndrome. *J Infect Dis* 2007; **196** Suppl 1:S63–S75.
- 62 Njuguna C, Stewart A, Mouton JP, Blockman M, Maartens G, Swart A, *et al.* Adverse drug reactions reported to a national hiv & tuberculosis health. *Drug Saf* 2016; **39**:159–169.
- 63 McDonagh EM, Boukouvala S, Akiillu E, Hein DW, Altman RB, Klein TE PharmGKB summary: very important pharmacogene information for N-acetyltransferase 2. *Pharmacogenet Genomics* 2014; **24**:409–425.
- 64 Wilkins JJ, Langdon G, McIlleron H, Pillai G, Smith PJ, Simonsson US. Variability in the population pharmacokinetics of isoniazid in South African tuberculosis patients. *Br J Clin Pharmacol* 2011; **72**:51–62.
- 65 Wondwossen A, Waqtola C, Gameda A. Incidence of antituberculosis-drug-induced hepatotoxicity and associated risk factors among tuberculosis patients in Dawro Zone, South Ethiopia: a cohort study. *Int J Mycobacteriol* 2016; **5**:14–20.
- 66 Ben Mahmoud L, Ghozzi H, Kamoun A, Hakim A, Hachicha H, Hammami S, *et al.* Polymorphism of the N-acetyltransferase 2 gene as a susceptibility risk factor for antituberculosis drug-induced hepatotoxicity in Tunisian patients with tuberculosis. *Pathol Biol (Paris)* 2012; **60**:324–330.
- 67 Matimba A, Oluka MN, Ebeshi BU, Sayi J, Bolaji OO, Guantai AN, *et al.* Establishment of a biobank and pharmacogenetics database of African populations. *Eur J Hum Genet* 2008; **16**:780–783.
- 68 Roederer MW, McLeod H, Juliano JJ. Can pharmacogenomics improve malaria drug policy? *Bull World Health Organ* 2011; **89**:838–845.
- 69 Kerb R, Fux R, Mörike K, Kremsner PG, Gil JP, Gleiter CH, *et al.* Pharmacogenetics of antimalarial drugs: effect on metabolism and transport. *Lancet Infect Dis* 2009; **9**:760–774.
- 70 Scott SA, Sangkuhl K, Shuldiner AR, Hulot JS, Thorn CF, Altman RB, *et al.* PharmGKB summary: very important pharmacogene information for cytochrome P450, family 2, subfamily C, polypeptide 19. *Pharmacogenet Genomics* 2012; **22**:159–165.
- 71 Hicks JK, Bishop JR, Sangkuhl K, Muller DJ, Ji Y, Leckband SG, *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and CYP2C19 genotypes and dosing of selective serotonin reuptake inhibitors. *Clin Pharmacol Ther* 2015; **98**:127–134.
- 72 Scott SA, Sangkuhl K, Stein CM, Hulot JS, Mega JL, Roden DM, *et al.* Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. *Clin Pharmacol Ther* 2013; **94**:317–323.
- 73 Kaneko A, Bergqvist Y, Taleo G, Kobayakawa T, Ishizaki T, Bjorkman A. Proguanil disposition and toxicity in malaria patients from Vanuatu with high frequencies of CYP2C19 mutations. *Pharmacogenetics* 1999; **9**:317–326.
- 74 De Vos A, vdW J, Looovers HM. Association between CYP2C19\*17 and metabolism of amitriptyline, citalopram and clomipramine in Dutch hospitalized patients. *Pharmacogenomics J* 2011; **11**:359–367.
- 75 Motshoge T, Tawe L, Muthoga CW, Allotey J, Romano R, Quaye I, *et al.* Cytochrome P450 2C8\*2 allele in Botswana: human genetic diversity and public health implications. *Acta Trop* 2016; **157**:54–8.
- 76 Alessandrini M, Sahle A, Dodgen TM, Warnich L, Pepper MS. Cytochrome P450 pharmacogenetics in African populations. *Drug Metab Rev* 2013; **45**:253–275.
- 77 Burrows JN, Duparc S, Gutteridge WE, Huijsduijnen RHV, Kaszubska W, Macintyre F, *et al.* New developments in anti-malarial target candidate and product profiles. *Malar J* 2017; **16**:26.
- 78 Cavallari LH, Langae TY, Momary KM, Shapiro NL, Nutescu EA, Coty WA, *et al.* Genetic and clinical predictors of warfarin dose requirements in African Americans. *Clin Pharmacol Ther* 2010; **87**:459–464.
- 79 Shahin MH, Khalifa SI, Gong Y, Hammad LN, Sallam MT, El Shafey M, *et al.* Genetic and nongenetic factors associated with warfarin dose requirements in Egyptian patients. *Pharmacogenet Genomics* 2011; **21**:130–135.
- 80 Johnson JA, Gong L, Whirl-Carrillo M, Gage BF, Scott SA, Stein CM, *et al.* Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. *Clin Pharmacol Ther* 2011; **90**:625–629.
- 81 Mitchell C, Gregersen N, Krause A. Novel CYP2C9 and VKORC1 gene variants associated with warfarin dosage variability in the South African black population. *Pharmacogenomics* 2011; **12**:953–963.
- 82 Krueger SK, Williams DE. Mammalian flavin-containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism. *Pharmacol Ther* 2005; **106**:357–387.