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.
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 include 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+ recovery rates [6].

The influence of variants on antimicrobial therapy has also received increased attention. Studies have reported associations between SLC01B1 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 coinfected 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 Khoean 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 56 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* list summarizing the genes investigated in this study according to biological function

<table>
<thead>
<tr>
<th>Category or function</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transporter</td>
<td>ABCC10, ABCB1, ABCC2, ABCC4, ABCG2, SLC1A5A, SLC22A1, SLC22A11, SLC22A6, SLC22A8, SLC35D1, SLC01A2, SLC01B1, SLC01B3, SLC02A1, SLC02B1, SLC03A1, SLC04A1, SLC04C1</td>
</tr>
<tr>
<td>Drug-metabolizing enzyme</td>
<td>CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C18, CYP2C8, CYP2C9, CYP2D6, CYP2E1, CYP3A4, CYP3A5, CYP3A7, CYP5B1, DP9D, ANPERP, GGT1, GSTM1, UGT1A1, CES1, CES2, FMO1, FMO2, FMO3, NAT1, NAT2, SULT1A1</td>
</tr>
<tr>
<td>Renal function</td>
<td>B2M, CLU, LCN2, TIMP1, HAVCR1 (or KIM)</td>
</tr>
<tr>
<td>Immune function</td>
<td>HLAB, HLAC, HLADR81, IL18</td>
</tr>
<tr>
<td>Transcription factor</td>
<td>AHR, NR112, NR113</td>
</tr>
<tr>
<td>Other</td>
<td>AGTR2, AGTRAP, ACE, FGA, FGG, TFF3, POR, NAGA</td>
</tr>
</tbody>
</table>

*Complete details of the ion AmpliSeq gene panel are available through the following link: https://www.ampliseq.com/protected/shared/viewSharedDesign.action?key=809e12aa81cb58e8caaa06e417c9f6c2cb500ea1966ad1332d6fde11c91f3c2b1.

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 homoyzogous 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/
B355) from 1996 variants originally identified. In addition, three variants more than 30bp in size were observed, including a 32bp deletion in POR, a 65bp deletion in HAVCR1 and a 51bp 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). 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).

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
Targeted ultra-deep sequencing of a South African Bantu-speaking cohort Tshabalala et al. 171

(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 Wolayta 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

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Variant genome coordinates</th>
<th>dbSNP rs number</th>
<th>Reference allele</th>
<th>Alternate allele</th>
<th>Gene region</th>
<th>Consequence</th>
<th>Protein modificationa</th>
<th>Bantu cohort frequency of alternate allele (global frequencyb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGTRAP</td>
<td>1:11808597</td>
<td>rs771467011</td>
<td>C</td>
<td>T</td>
<td>Exon</td>
<td>Stop gained</td>
<td>p.Leu98=</td>
<td>0.0375 (0.0008)</td>
</tr>
<tr>
<td>AGTRAP</td>
<td>1:11810271</td>
<td>rs531367893</td>
<td>–</td>
<td>CCCCTGCC</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>–</td>
<td>0.0125 (0.0048)</td>
</tr>
<tr>
<td>FM02</td>
<td>1:171165803</td>
<td>rs28368860</td>
<td>G</td>
<td>–</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Val113Ter</td>
<td>0.0225 (0.1028)</td>
</tr>
<tr>
<td>FM02</td>
<td>1:171171321</td>
<td>rs2020886</td>
<td>C</td>
<td>T</td>
<td>Exon</td>
<td>Stop gained</td>
<td>p.Arg249Ter</td>
<td>0.0125 (0.0048)</td>
</tr>
<tr>
<td>FM02</td>
<td>1:171178090</td>
<td>rs6881174</td>
<td>T</td>
<td>C</td>
<td>Exon</td>
<td>Stop lost</td>
<td>p.Ter472GlnexTer64</td>
<td>0.2125 (0.0463)</td>
</tr>
<tr>
<td>FM01</td>
<td>1:171254588</td>
<td>rs60639054</td>
<td>C</td>
<td>T</td>
<td>Exon</td>
<td>Stop gained</td>
<td>p.Arg502Ter</td>
<td>0.0112 (0.0040)</td>
</tr>
<tr>
<td>SLC15A2</td>
<td>3:121613361</td>
<td>–</td>
<td>–</td>
<td>T</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Ser15Phefs</td>
<td>0.0125 (NA)</td>
</tr>
<tr>
<td>HAVCR1</td>
<td>5:156456731</td>
<td>rs111756551</td>
<td>G</td>
<td>T</td>
<td>Exon</td>
<td>Stop gained</td>
<td>p.Cys358Ter</td>
<td>0.0225 (0.0114)</td>
</tr>
<tr>
<td>HLA-C</td>
<td>6:31238001</td>
<td>–</td>
<td>–</td>
<td>C</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Thr295Hisfs</td>
<td>0.025 (NA)</td>
</tr>
<tr>
<td>HLA-B</td>
<td>6:31323953</td>
<td>–</td>
<td>–</td>
<td>G</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Glu204Alafs</td>
<td>0.075 (NA)</td>
</tr>
<tr>
<td>HLA-B</td>
<td>6:31324176</td>
<td>–</td>
<td>C</td>
<td>–</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Pro129=</td>
<td>0.0125 (NA)</td>
</tr>
<tr>
<td>SLC22A1</td>
<td>6:16054329</td>
<td>rs111863691</td>
<td>–</td>
<td>G</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Cys121Trpfs</td>
<td>0.0125 (NA)</td>
</tr>
<tr>
<td>SLC22A1</td>
<td>6:160560898</td>
<td>rs113569197</td>
<td>TGGTAAAGT</td>
<td>–</td>
<td>Exon</td>
<td>Splice donor variant</td>
<td>–</td>
<td>0.3375 (0.6895)</td>
</tr>
<tr>
<td>CYP3A5</td>
<td>7:99250394</td>
<td>rs41303343</td>
<td>A</td>
<td>–</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Thr346Tyrfs</td>
<td>0.0875 (0.0315)</td>
</tr>
<tr>
<td>CYP3A5</td>
<td>7:99250394</td>
<td>rs8977185</td>
<td>T</td>
<td>C</td>
<td>Exon</td>
<td>Stop lost</td>
<td>–</td>
<td>0.1 (0.0649)</td>
</tr>
<tr>
<td>CYP2C8</td>
<td>10:96827210</td>
<td>rs1182072</td>
<td>–</td>
<td>A</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>–</td>
<td>0.2 (0.3498)</td>
</tr>
<tr>
<td>CYP2C8</td>
<td>10:9569915</td>
<td>rs142470035</td>
<td>T</td>
<td>C</td>
<td>Exon</td>
<td>Start lost</td>
<td>p.Met1?</td>
<td>0.0125 (0.0081)</td>
</tr>
<tr>
<td>ABCC2</td>
<td>10:10157712</td>
<td>TT</td>
<td>–</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Phe721Trpfs</td>
<td>0.0125 (NA)</td>
<td></td>
</tr>
<tr>
<td>CES1</td>
<td>16:53866915</td>
<td>rs139063675</td>
<td>C</td>
<td>T</td>
<td>Intron</td>
<td>Splice donor variant</td>
<td>–</td>
<td>0.0375 (0.0008)</td>
</tr>
<tr>
<td>GSTT1</td>
<td>22:25023506</td>
<td>–</td>
<td>G</td>
<td>–</td>
<td>Exon</td>
<td>Frameshift variant</td>
<td>p.Asp377Metfs</td>
<td>0.0125 (NA)</td>
</tr>
<tr>
<td>NAGA</td>
<td>22:24246173</td>
<td>CATCCTGGTA</td>
<td>–</td>
<td>Intron</td>
<td>Splice donor variant</td>
<td>–</td>
<td>0.0125 (NA)</td>
<td></td>
</tr>
<tr>
<td>CYP2D6</td>
<td>22:24252494</td>
<td>m30892097</td>
<td>C</td>
<td>T</td>
<td>Intron</td>
<td>Splice acceptor variant</td>
<td>–</td>
<td>0.0375 (0.0093)</td>
</tr>
</tbody>
</table>

aKey: fs, frameshift; Ter, termination.

bGlobal alternate allele frequencies are based on data from the 1000 Genomes Project phase 3.
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 10 kb blocks along the genome identified regions in chromosomes 1 (DYPD), 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 ABCC1 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 ≥ 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 SLC01B1*5 (rs4149056)], highlighting their potential clinical impact. For example, a novel, moderately frequent (MAF = 0.075) missense SLC03A1 variant (15:92694224 T/C) was assigned a deleterious (SIFT score 0) variant (PolyPhen 2.0 score 0.99) phenotype, and a rarer (MAF = 0.013), novel missense variant (2234669414 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 10 kb 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 (DYPD), chromosome 6 (HLA–C) and 13 (ABCC4). These genes play important roles in pyrimidine metabolism, HIV

<table>
<thead>
<tr>
<th>Variant prediction tool</th>
<th>Prediction</th>
<th>Number of variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>PolyPhen 2.0</td>
<td>Possibly damaging (0.8–0.9)</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Probably damaging (≥ 0.9)</td>
<td>46</td>
</tr>
<tr>
<td>SIFT</td>
<td>Deleterious (≤ 0.05)</td>
<td>97</td>
</tr>
<tr>
<td>Consensus of PolyPhen</td>
<td>Possibly damaging and deleterious</td>
<td>17</td>
</tr>
<tr>
<td>and SIFT</td>
<td>Deleterious</td>
<td>41</td>
</tr>
</tbody>
</table>

Table 3 Functional analysis of variants identified in the cohort
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.

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-CEA (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.

![Heat map illustrating allele frequencies](image)

**Fig. 3**

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), a Chinese (CHB) and a European (CEU) population. Frequencies are depicted across a shade spectrum ranging from relatively rare (blue) to relatively frequent (red).

**Fig. 4**

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>SNP ID</th>
<th>GENE</th>
<th>STUDY COHORT</th>
<th>AGVP_ZUL</th>
<th>KGP_YRI</th>
<th>KGP_LWK</th>
<th>KGP_CHEB</th>
<th>KGP_CEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug Transport</td>
<td>rs5995889</td>
<td>ABC4C</td>
<td>0.431</td>
<td>0.295</td>
<td>0.181</td>
<td>0.247</td>
<td>0.228</td>
<td>0.151</td>
</tr>
<tr>
<td>Drug Metabolism (Poor)</td>
<td>rs1440917</td>
<td>SLC16A3</td>
<td>0.347</td>
<td>0.255</td>
<td>0.247</td>
<td>0.278</td>
<td>0.401</td>
<td>0.136</td>
</tr>
<tr>
<td>Drug Metabolism (Intermediate)</td>
<td>rs3734274 (SOLCO1B1)</td>
<td>ABC2</td>
<td>0.444</td>
<td>0.350</td>
<td>0.409</td>
<td>0.445</td>
<td>0.283</td>
<td>0.631</td>
</tr>
<tr>
<td>Drug Metabolism</td>
<td>rs655986 (SOLCO1B1)</td>
<td>ABC2</td>
<td>0.181</td>
<td>0.170</td>
<td>0.238</td>
<td>0.192</td>
<td>0.136</td>
<td>0.237</td>
</tr>
<tr>
<td>Acetylators (Slow)</td>
<td>rs1002085 (NAT2)</td>
<td>0.422</td>
<td>0.090</td>
<td>0.116</td>
<td>0.091</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Acetylators (Rapid)</td>
<td>rs12151419 (NAT2)</td>
<td>0.216</td>
<td>0.170</td>
<td>0.150</td>
<td>0.238</td>
<td>0.136</td>
<td>0.237</td>
<td></td>
</tr>
</tbody>
</table>

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 to other African populations, Chinese and European populations.
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 *ABCG2* 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 coinfected 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 Ameur 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 # V6YETS0) and a CSIR parliametary (grant # VIYBT’96) (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


178 Pharmacogenetics and Genomics 2019, Vol 29 No 7


